The University Of Sheffield.

Access To Thesis.

- A bound copy of every thesis which is accepted as worthy for a higher degree, must be deposited in the University of Sheffield Library, where it will be made available for borrowing or consultation in accordance with University Regulations.
- All students registering from 2008–09 onwards are also required to submit an electronic copy of their final, approved thesis. Students who registered prior to 2008–09 may also submit electronically, but this is not required.

Author: Miss Nuttawan PRAMANPOL Dept: Molecular Biology & Biotechnology

Thesis Title: Structural studies on immunogenic proteins of Burkholderia pseudomallei Registration No: 080125549

For completion by all students:

Submit in print form only (for deposit in the University Library): ☐

Submit in print form and also upload to the *White Rose eTheses Online* server: In full ☑ Edited eThesis ☐

Please indicate if there are any embargo restrictions on this thesis. Please note that if no boxes are ticked, you will have consented to your thesis being made available without any restrictions.

Embargo details: (complete only if requesting an embargo to either your print and/or eThesis)	Embargo required?		Length of embargo (in years)
Print Thesis	Yes ☐	No ☑	
eThesis	Yes ☑	No ☐	2

Supervisor: I, the supervisor, agree to the named thesis being made available under the conditions specified above.

Name: PATRICK BAKER Dept: MBB

Signed: [signature] Date: 22.3.13

Student: I, the author, agree to the named thesis being made available under the conditions specified above.

I give permission to the University of Sheffield to reproduce the print thesis in whole or in part in order to supply single copies for the purpose of research or private study for a non-commercial purpose.

I confirm that this thesis is my own work, and where materials owned by a third party have been used copyright clearance has been obtained. I am aware of the University's *Guidance on the Use of Unfair Means* (www.sheffield.ac.uk/lets/design/unfair)

I confirm that all copies of the thesis submitted to the University (including electronic copies on CD/DVD) are identical in content.

Name: Miss Nuttawan PRAMANPOL Dept: Molecular Biology & Biotechnology

Signed: Nuttawan Pramanpol Date: 22/3/13

For completion by students also submitting an electronic thesis (eThesis):

I, the author, agree that the University of Sheffield's eThesis repository (currently WREO) will make my eThesis available over the internet via an entirely non-exclusive agreement and that, without changing content, WREO may convert my thesis to any medium or format for the purpose of future preservation and accessibility.

I, the author, agree that the metadata relating to the eThesis will normally appear on both the University's eThesis server and the British Library's EThOS service, even if the thesis is subject to an embargo. I agree that a copy of the eThesis may be supplied to the British Library.

I confirm that the upload is identical to the final, examined and awarded version of the thesis as submitted in print to the University for deposit in the Library (unless edited as indicated above).

Name: Miss Nuttawan PRAMANPOL Dept: Molecular Biology & Biotechnology

Signed: Nuttawan Pramanpol Date: 22/3/13

THIS SHEET MUST BE BOUND IN THE FRONT OF THE PRINTED THESIS BEFORE IT IS SUBMITTED

Structural Studies on Immunogenic Proteins of *Burkholderia pseudomallei*

Nuttawan Pramanpol

A thesis submitted to the University of Sheffield in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Molecular Biology and Biotechnology
University of Sheffield

March 2013

Abstract

Burkholderia pseudomallei is a Gram negative bacterial pathogen and the causative agent of melioidosis, which is the third most frequent cause of death from infectious disease in Northeastern Thailand and also highly endemic in Southeast Asia and Northern Australia. Problems of melioidosis management include misdiagnosis and the increasing resistance of the bacteria to the current anti-infective treatment, resulting in high mortality and having a global impact as a potential bioweapon. Little is known about the ambiguous pathogenesis and virulence mechanisms of *B. pseudomallei* infection. One way to gain understanding is through the investigation of the structure/function relationships of potential target proteins, as a starting point for a rational drug discovery programme. Structural studies on the target proteins of *B. pseudomallei* can shed light on their biological functions at the molecular level.

This thesis presents the structural studies on immunogenic proteins of *B. pseudomallei* as part of a targeted structural genomics programme on *B. pseudomallei*. Seven target genes were selected from lists of putative essential genes and also identified *B. pseudomallei* proteins that elicit an immuno cross reactivity to the sera of infected humans, which may be involved in virulence and/or pathogenicity of *B. pseudomallei*. Of those, BPSL0606, an uncharacterised protein, was cloned, over-expressed, purified, and crystallised. The first crystal structure of BPSL0606 was determined and revealed that it is structurally similar to the GCN5-related N-acetyltransferase (GNAT) superfamily. Dimer formation in the crystal packing agreed with gel filtration analysis and the interface of BPSL0606 was similar to many GNAT members. Structure-based alignment of BPSL0606 against known structures of GNATs showed that BPSL0606 has conserved residues that are involved in the binding of acetyl coenzyme A in members of GNATs, suggesting that BPSL0606 may bind the acetyl coenzyme A. Although co-crystallization of BPSL0606 and acetyl coenzyme A and ligand observation by nuclear magnetic resonance suggested that BPSL0606 did not bind to the acetyl coenzyme A. Analysis of the aminoglycoside binding site indicated that BPSL0606 could possibly bind to this substrate. The BPSL0606 structure with acetyl coenzyme A and kanamycin was modelled, proposing a possible clue related to the antibiotic resistance. The findings in this PhD thesis provide the molecular and structural information of *B. pseudomallei* proteins, which could contribute to further investigations.

Acknowledgements

During the PhD study, I have been supported by many people around the world. Without their encouragements, my PhD journey would not have been completed.

I would like to thank my supervisor, Dr. Patrick J. Baker for giving me an opportunity to pursue a doctoral degree. He has taught, trained and guided me with his patience and politeness throughout the four-year time of study. He always encourages me when things go wrong and supports me to gain experience and develop skills in such great opportunities, for example, attending WMC in Australia, visiting UKM in Malaysia and attending PX summer school, EMBO practical course and annual CCP4 study weekend in UK. All of these really take part of my academic achievement.

I would like to acknowledge Dr. Svetlana Sedelnikova, who advised me in protein purification from her technical expertise. I would also like to give thanks to my advisors, Prof. David Rice and Dr. Rosie Staniforth for their advices and comments on my PhD research project. I am indebted Dr. Jeremy Craven and Andrea Hounslow for their help in NMR experiments.

I really appreciate helpfulness and friendliness from people in the crystallography group, including Prof. Pete Artymiuk, Dr. John Rafferty, Fiona Rodgers, Dr. Linda Britton, Dr. Claudine Bisson, Dr. Fabia Allen, Dr. Chatrudee Suwannachart, Dr. Zaza Ramly, Dr. Abimael Cruz-Migoni, Ghani Abd Aziz, Magdah Ganash and all others.

I am truly thankful to Malaysian collaborators and visitors who have provided materials for this study and shared and exchanged their expertise and experience.

I would like to give special thanks to Sawaporn Siripanthana for being with me and sharing the whole journey of my PhD. Many thanks to Pimpunyawat Tummuangpak for wonderful friendship. I sincerely thank Suchittra Wongprayoon, who keeps giving me hope. I appreciate kindly support from Thais in Sheffield and friends in Thailand.

An invaluable love from my family has driven me pass through every single step from the beginning of my life. I am tremendously grateful having them, who always give me support and incentive, unconditionally.

Last but not least, I would like to acknowledge the Royal Thai Government for funding and support from National Synchrotron Research Institute, Ministry of Science and Technology, Thailand.

Table of Contents

Chapter 1: Introduction	**1**
1.1. Microbiological aspect of *B. pseudomallei*	2
1.2. Geographical distribution of melioidosis	3
1.3. Clinical presentation	4
1.4. *B. pseudomallei* genome	6
1.5. Pathogenesis and virulence	
1.5.1. Pathogenesis	7
1.5.2. Virulence determinants	8
1.6. Host-pathogen interaction	
1.6.1. Pathogen recognition	12
1.6.2. Bacterial intracellular survival	12
1.7. Melioidosis management	
1.7.1. Diagnosis	13
1.7.2. Antibiotic resistance	14
1.8. Identification of potential therapeutic target proteins	16
1.9. Structural genomics	17
1.10. Aims of the study and thesis outline	20
Chapter 2: Protein Crystallography	**21**
2.1. Protein crystallization	21
2.1.1. Vapor diffusion techniques	23
2.1.1.1. Sitting drop vapor diffusion	23
2.1.1.2. Hanging drop vapor diffusion	23
2.1.2. Protein crystals	
2.1.2.1. Crystal assembly	24
2.1.2.2. Crystal system	24
2.1.2.3. Mounting crystals	26
2.2. X-ray diffraction	26
2.2.1. Principles: Bragg's law	26
2.2.2. Fourier Transform	27
2.3. Data collection	29
2.3.1. Data collection strategies	31
2.3.2. Radiation damage	32
2.3.3. Tools for data collection	33

2.3.3.1. X-ray sources and optics 33
2.3.3.1.1. Rotating anode 33
2.3.3.1.2. Synchrotron 34
2.3.3.2. Detectors 34
2.3.3.3. Goniometer 35
2.3.3.4. Cryocooler 35
2.4. Data processing 36
2.4.1. Autoindexing 36
2.4.2. Integration 36
2.4.3. Data reduction 37
2.4.4. Determining data quality 38
2.5. Structure determination 38
2.5.1. Patterson function 39
2.5.2. Isomorphous replacement 40
2.5.3. Multi-wavelength anomalous dispersion (MAD) 42
2.5.4. Molecular replacement 46
2.6. Structure refinement 46
2.6.1. Maximum likelihood 47
2.7. Structure validation 49

Chapter 3: Materials and Methods 51
3.1. Target selection 51
3.2. DNA manipulation
3.2.1. Gene amplification 51
3.2.2. Agarose gel electrophoresis 52
3.2.3. PCR purification and gel extraction 53
3.2.4. Gene cloning
3.2.4.1. TA cloning 53
3.2.4.2. Restriction enzyme cloning 54
3.2.5. Transformation of *E. coli* competent cells 55
3.2.6. Screening of recombinant clones by colony PCR 56
3.2.7. Plasmid DNA extraction 56
3.2.8. Analysis of DNA sequencing 57
3.3. Protein expression
3.3.1. Transformation of Tuner (DE3) pLacI 57
3.3.2. Transformation of BL21 (DE3) 57

3.3.3. Protein over-expression 58
3.3.4. Selenium-methionine (Se-Met) incorporated protein production 59
3.3.5. Protein expression analysis 60
3.4. Protein purification, crystallisation and data collection 62
3.5. Data processing 63
3.6. Structure determination 63
3.7. Ligand observation by nuclear magnetic resonance (NMR) 63

Chapter 4: Target Selection, Cloning, Protein Expression and Purification 65
4.1. Target selection 65
4.2. Gene amplification
4.2.1. Primer design 69
4.2.2 Polymerase chain reaction (PCR) 71
4.3. BPSL3022-Cell division protein FtsQ 73
4.4. BPSS0945-Subfamily M23B unassigned peptidase 85
4.5. BPSS0238 and BPSS1416
4.5.1. BPSS0238 Penicillin-binding protein 97
4.5.2. BPSS1416 Uncharacterized protein 100
4.6. Other uncharacterized proteins
4.6.1. BPSS0603 103
4.6.2. BPSS0683 107
4.6.3. BPSL0606 110
4.7. Chapter summary 118

Chapter 5: Crystallization, Analysis of X-ray Diffraction and Structure Determination of BPSL0606 121
5.1. Crystallization of BPSL0606 121
5.2. X-ray data collection 124
5.3. Data processing 124
5.4. Analysis of self-rotation function and self-Patterson function 127
5.5. Structure determination 128
5.6. Se-methionine protein 131
5.7. Structure validation 138
5.8. Structure description 141

5.9. Crystallization and structure determination of the truncated BPSL0606 145
5.9.1. Data collection and processing 147
5.9.2. Structure determination 148
5.9.3. Structure description 153

Chapter 6: Structure of BPSL0606 155
6.1. The crystal structure of BPSL0606 contained 4 molecules in the asymmetric unit ($P2_1$ crystal form) 155
6.1.1. Molecular surfaces 157
6.1.2. Dimer interfaces and assemblies 158
6.2. The crystal structure of the truncated BPSL0606 contained 2 molecules in the asymmetric unit ($P2_12_12_1$ crystal form) 163
6.3. Interface similarity 167

Chapter 7: Structural and functional analysis of BPSL0606 173
7.1. Structure similarity 173
7.2. Structural fold and motifs of GNAT family members 176
7.3. Structure comparison 180
7.4. Cocrystallization of BPSL0606 and acetyl coenzyme A 184
7.5. Ligand observation by NMR 190
7.6. Analysis of acetyl coenzyme A binding site 193
7.7. Analysis of substrate binding of GNATs 201
7.8. Chapter summary 206

Chapter 8: Conclusions and Future work 207
8.1. Summary 207
8.2 Future work 210

Appendix 213

References 215

List of Figures

Chapter 1

Figure 1.1 The variation of colony morphotypes of *B. pseudomallei* grown on Ashdown's selective media. 2

Figure 1.2 The worldwide distribution map of *B. pseudomallei* and melioidosis 4

Figure 1.3 Clinical presentations of melioidosis 5

Figure 1.4 Schematic diagrams of *B. pseudomallei* strain K96243 genome 6

Figure 1.5 The *B. pseudomallei* intracellular lifestyle 8

Figure 1.6 From structure to function 19

Chapter 2

Figure 2.1 Crystallization diagrams 22

Figure 2.2 Vapor diffusion experiments 24

Figure 2.3 The 14 Bravais lattices 25

Figure 2.4 Bragg's law 27

Figure 2.5 The Ewald construction 30

Figure 2.6 Isomorphous replacement 41

Figure 2.7 Typical absorption curve showing anomalous scattering signal for MAD phasing 42

Figure 2.8 Breakdown of Friedel's law 44

Figure 2.9 Argand diagrams involved in MAD phasing 45

Chapter 4

Figure 4.1 An initial target selection 67

Figure 4.2 Crystal structures of periplasmic FtsQ in *E.coli* and *Y.entericolitica* 73

Figure 4.3 A hydropathy plots for BPSL3022 74

Figure 4.4 Transmembrane prediction of BPSL3022 75

Figure 4.5 Protein sequence alignment of crystal structures of periplasmic FtsQ in *E. coli* and *Y. enterocolitica* and BPSL3022 76

Figure 4.6 Protein sequence analysis of BPSL3022 by ProtParam 77

Figure 4.7. PCR products of truncated BPSL3022 gene amplification on a 1% agarose gel 78
Figure 4.8 DNA sequencing of BPSL3022_2 in the pETBlue-1 clone 79
Figure 4.9 SDS PAGE BPSL3022_2 in pETBlue-1 expressed in *E.coli* Tuner (DE3) pLacI cells 80
Figure 4.10 DNA sequencing of the truncated FtsQ with C-terminal His tag clone (BPSL3022_3) 82
Figure 4.11 SDS PAGE of BPSL3022_3 in pET24-a expressed in *E.coli* BL21 (DE3). 83
Figure 4.12 Analysis of BPSS0945 protein sequence 85
Figure 4.13 A hydropathy plot of BPSS0945 86
Figure 4.14 Signal peptide prediction of BPSS0945 87
Figure 4.15 Protein sequence alignment of BPSS0945 and proteins with known structure 89
Figure 4.16 Amplification of BPSS0945 coding gene 90
Figure 4.17 DNA sequencing of the BPSS0945 recombinant clone 91
Figure 4.18 Overexpression of BPSS0945 92
Figure 4.19 Purification of BPSS0945 94
Figure 4.20 Stability of BPSS0945 94
Figure 4.21 BPSS0945 treated with EDTA 96
Figure 4.22 Analysis of BPSS0238 sequence 97
Figure 4.23 A hydropathy plot of BPSS0238 98
Figure 4.24 PCR product of BPSS0238 98
Figure 4.25 Analysis of BPSS1416 sequence 100
Figure 4.26 A hydropathy plot of BPSS1416 101
Figure 4.27 PCR product of BPSS1416 102
Figure 4.28 Analysis of BPSS0603 sequence 103
Figure 4.29 A hydropathy plot of BPSS0603 104
Figure 4.30 PCR product of BPSS0603 amplification 105
Figure 4.31 DNA sequencing of the recombinant BPSS0603 clone 105

Figure 4.32 Overexpression of BPSS0603 106

Figure 4.33 ProtParam of BPSS0683 107

Figure 4.34 A hydropathy plots of BPSS0683 108

Figure 4.35 PCR product of BPSS0683 amplification 108

Figure 4.36 Overexpression of BPSS0683 109

Figure 4.37 The characterization of BPSL0606 sequence by ProtParam 110

Figure 4.38 Schematic of BPSL0606 domain 111

Figure 4.39 Kyte-Doolittle hydropathy plot of BPSL0606 111

Figure 4.40 Analysis of PCR product of BPSL0606 112

Figure 4.41 DNA sequencing of the recombinant BPSL0606 clone 113

Figure 4.42 BPSL0606 expression profile 114

Figure 4.43 Analysis of BPSL0606 protein purification 116

Figure 4.44 Analysis of purified BPSL0606 by mass spectrometry 117

Chapter 5

Figure 5.1 Crystals of BPSL0606 122

Figure 5.2 Data collection of the native BPSL0606 crystal 123

Figure 5.3 Analysis of self-rotation function 127

Figure 5.4 Fold recognition and secondary structure prediction of BPSL0606 128

Figure 5.5 Crystal packing for 4 copies of BPSL0606 129

Figure 5.6 The C-terminal 65 residues of BPSL0606 were missing 130

Figure 5.7 Biased structure determination by molecular replacement 131

Figure 5.8 Crystals of the Se-Met protein of BPSL0606 131

Figure 5.9 A fluorescence absorption spectrum of the Se-Met BPSL0606 crystal near to the absorption K-edge of Se 132

Figure 5.10 MAD data collection 133

Figure 5.11 Selenium atom location 136

Figure 5.12 A portion of the electron density map of BPSL0606 137

Figure 5.13 Ramachandran plots for the BPSL0606 structure 139

Figure 5.14 Cartoon representation of the secondary structure of BPSL0606 141

Figure 5.15 Protein disorder prediction 142

Figure 5.16 Available space in the crystal lattice beyond residue 151 143

Figure 5.17 SDS PAGE of the BPSL0606 crystals 144

Figure 5.18 Mass spectrometry of the remaining purified protein 144

Figure 5.19 Mass spectrometry of the old protein used for the crystallization experiment 145

Figure 5.20 A cryoprotected crystal of the truncated BPSL0606 146

Figure 5.21 The final model of the truncated BPSL0606 149

Figure 5.22: Example of other molecules found in the crystal structure 149

Figure 5.23 Ramachandran plots for the truncated BPSL0606 structure 151

Chapter 6

Figure 6.1 Cartoon representation of the crystal structure of BPSL0606 155

Figure 6.2 An overlap of 4 subunits of BPSL0606 structure 157

Figure 6.3 Superposition of the A-B and C-D dimers 159

Figure 6.4 Interfacing residues of the dimer A-B 161

Figure 6.5 Cartoon representation of the crystal structure of the truncated BPSL0606 163

Figure 6.6 An overlap of 2 subunits in the truncated BPSL0606 structure 164

Figure 6.7 Crystal packing of BPSL0606 in two crystal forms 165

Figure 6.8 A similar dimer interface to BPSL0606 169

Figure 6.9 A similar dimer interface to BPSL0606 but different orientation 170

Figure 6.10 Different oligomeric assemblies with similar interface to BPSL0606 170

Figure 6.11 Different dimer interfaces to BPSL0606 171

Chapter 7

Figure 7.1 Superposition of BPSL0606 and similar structures 175

Figure 7.2 Schematic of the general reaction of GNATs catalysis 177

Figure 7.3 The topology of the GNAT fold 179

Figure 7.4 The acetyl coenzyme A binding site in the β4-α3-β5 motif 179

Figure 7.5 Superposition of the BPSL0606 structure and GNAT structures 180

Figure 7.6 Structure-based alignment of BPSL0606 and GNAT proteins from different species 181

Figure 7.7 Structure based sequence alignment among the GNAT members and BPSL0606 182

Figure 7.8 Conserved residues of BPSL0606 183

Figure 7.9 No electron density for acetyl coenzyme A 187

Figure 7.10 Still no acetyl coenzyme A bound to BPSL0606 189

Figure 7.11 An example of ligand binding observation 190

Figure 7.12 1D ^{1}H NMR spectra at 600 MHz for ligand observation 191

Figure 7.13 1D ^{1}H NMR spectra at 600 MHz of the acetyl coenzyme A solution 192

Figure 7.14 Superposition of the acetyl coenzyme A bound GNAT structures 193

Figure 7.15 Different conformation of bound acetyl coenzyme A 195

Figure 7.16 Superposition of the BPSL0606 and the aminoglycoside 6'-N-acetyltransferase from *E. faecium* (PDB ID: 1B87) 199

Figure 7.17 The conformational change of serotonin N-acetyltransferase 200

Figure 7.18 Crystal structure of aminoglycoside 6' N-acetyltransferase from *E. coli* complexed with acetyl coenzyme A and different aminoglycoside substrates 203

Figure 7.19 The active site of AAC6' 204

Figure 7.20 Surface electrostatics of the structures of AAC6' complex with substrates and BPSL0606 204

Figure 7.21 The proposed kanamycin binding site binding site of BPSL0606 205

Figure 7.22 The proposed acetyl coenzyme A binding site of BPSL0606 205

Figure 7.23 Model of BPSL0606 complex with acetyl coenzyme A and kanamycin 206

Chapter 8

Figure 8.1 Pyramid of attrition rate in key steps of structure determination project 208

Appendix	
Figure 3.1 pETBlue-1 vector map and cloning and expression region	213
Figure 3.2 pET24-a vector map and cloning and expression region	214

List of Tables

Chapter 3

Table 3.1 Recipe for PCR 52

Table 3.2 A protocol of ligation using pETBlue-1™ AccepTor vector 53

Table 3.3 Restriction enzyme digestion 54

Table 3.4 Ligation for recombinant pET24-a 55

Table 3.5 Lysogeny broth (LB) media and LB agar recipes 56

Table 3.6 Minimal media preparation for Se-Met substituted protein production 59

Table 3.7 SDS polyacrylamide gel recipe 62

Table 3.8 10X SDS buffer recipe 62

Chapter 4

Table 4.1 A list of selected targets 69

Table 4.2 A list of primers used in this work 70

Table 4.3 PCR conditions for target gene amplification 72

Table 4.4 A list of similar protein structures to the predicted BPSS0945 structure 88

Table 4.5 Summary of target proteins studied in this project. 118

Chapter 5

Table 5.1 Data collection statistics of the native BPSL0606 crystal 126

Table 5.2 Data collection statistics of the Se-Met protein crystal of BPSL0606 134

Table 5.3 Statistics of the selenium sites identified by experimental phasing 135

Table 5.4 Data refinement statistics of BPSL0606 structure model 138

Table 5.5 Data collection statistics of the truncated BPSL0606 crystal 147

Table 5.6 Data refinement statistics of the truncated BPSL0606 150

Chapter 6

Table 6.1 Analysis of XYZ displacement between 4 subunits in the BPSL0606 structure by LSQKAB 156

Table 6.2 Analysis of surface areas of each subunit in the BPSL0606 structure 158

Table 6.3 Analysis of dimer interface in the BPSL0606 structure 158

Table 6.4 Analysis of A-B dimer interfacing residues of the BPSL0606 structure 160

Table 6.5 Interaction of H-bond between Chain A and B of the BPSL0606 structure 160

Table 6.6 Analysis of C-D dimer interfacing residues of the BPSL0606 structure 162

Table 6.7 Interaction of H-bond between Chain C and D of the BPSL0606 structure 162

Table 6.8 Analysis of surface areas of each subunit in the crystal form $P2_12_12_1$ 166

Table 6.9 Analysis of A-B dimer interfacing residues the crystal form $P2_12_12_1$ 166

Table 6.10 Interaction of H-bond between Chain A and B the crystal form $P2_12_12_1$ 167

Table 6.11 Interface similarity of the dimer in the crystal form $P2_1$ 167

Table 6.12 Interface similarity of the dimer in the crystal form $P2_12_12_1$ 168

Chapter 7

Table 7.1 Top-ten hits for matching the BPSL0606 structure by Dali server 174

Table 7.2 Crystallization conditions for the relative GNAT proteins 185

Table 7.3 Data collection statistics of crystals from co-crystallization experiments 186

Table 7.4 Data collection statistics of crystals from soaking experiments 188

Table 7.5 The main chain interaction of the acetyl coenzyme A bound to AAC6' 198

Table 7.6 Comparison of residues involved in side chain interaction made between AAC6' and acetyl coenzyme A and equivalent residues in BPSL0606 199

Abbreviations and Symbols

Crystallographic terms

Å	Angstrom
a,b,c	Real space unit cell dimensions
α,β,γ	Real space unit cell angles
hkl	Miller indices
d_{hkl}	Inter-planar spacing between a set of parallel planes
λ	Wavelength
θ	Diffraction angle
x,y,z	Real space coordinates
$\mathbf{F}_{hkl}$	Structure factor of a reflection hkl
$\lvert\mathbf{F}_{hkl}\rvert$	Structure factor amplitude
ρ (x,y,z)	Electron density at a position (x,y,z)
α_{hkl}	Phase angle
I_{hkl}	Intensity of a reflection hkl
R_{free}	Free R factor
R_{merge}	Merging R factor
R_{pim}	Precision indicating merging R-factor
$I/\sigma(I)$	Signal to noise ratio
P (u,v,w)	Patterson function
$\mathbf{F}_P$	Structure factor for a protein structure
$\mathbf{F}_H$	Structure factor for a heavy atom substructure
$\mathbf{F}_{PH}$	Structure factor for a heavy atom derivative structure
f	An atomic scattering factor
f^0	Normal scattering contribution
f'	Dispersive scattering contribution
f''	Anomalous scattering contribution
$\mathbf{F}_T$	Structure factor for a protein without anomalous
$\mathbf{F}_A$	Structure factor for an anomalous marker
$\lvert\mathbf{F}^+\rvert$, $\lvert\mathbf{F}^-\rvert$	Structure factor amplitudes of reflections related by Friedel's law
φ_T	Protein phase
φ_A	Anomalous scattering contribution phase

$\|\mathbf{F}_{obs}\|$	Observed structure factor amplitude in reciprocal space
$\|\mathbf{F}_{cal}\|$	Calculated structure factor amplitude in real space
LL	Log-likelihood
LLG	Log-likelihood gain
RMSD	Root mean square deviation
MAD	Multi wavelength anomalous dispersion

Biological and chemical terms

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
PCR	Polymerase chain reaction
dNTP	Deoxyribonucleotide triphosphate
A	Adenine
T	Thymine
G	Guanine
C	Cytosine
DMSO	Dimethyl sulphoxide
IPTG	Isopropyl-β-D-thiogalactoside
X-gal	5-Bromo-4-chloro-3-indolyl- β-D-galactoside
HCl	Hydrochloric acid
Na_2HPO_4	Disodium hydrogenphosphate
KH_2PO_4	Potassium dihydrogenphosphate
NaCl	Sodium chloride
NH_4Cl	Ammonium chloride
Se-Met	Seleno-methionine
TEMED	Tetramethylethylenediamine
EDTA	Ethylenediamine tetraacetic acid
SDS	Sodium dodecyl sulphate
PAGE	Polyacrylamide gel electrophoresis
PEG	Polyethylene glycol
PDB	Protein data bank
H-bond	Hydrogen bond
GNAT	GCN-5 related N-acetyltransferase
AAC	Aminoglycoside N-acetyltransferase

ADP	Adenosine diphosphate
UV	Ultraviolet
NMR	Nuclear magnetic resonance

Miscellaneous

OD_{600}	Optical density at a wavelength of 600 nm
A_{595}	Absorbance at a wavelength of 595 nm
bp	Base pair
Da	Dalton
pI	Isoelectric point
MW	Molecular weight
MWCO	Molecular weight cut off
rpm	rotation per minute

Units of measurement are expressed in SI.

Amino acid residues are in single or three letter codes.

Chapter 1: Introduction

Despite the prevalence of antibiotics and immunisation regimes, infectious diseases retain their ability to kill and thus understanding their pathogenicity has important implications for human health. Melioidosis is one such infectious disease and is endemic in Southeast Asia and northern Australia (Currie *et al*, 2008). The mortality of melioidosis is about 50% and 19% in northeast Thailand and northern Australia, respectively (Peacock, 2006). In northeast Thailand, melioidosis is currently the third most common cause of death from infectious diseases after HIV/AIDS and tuberculosis (Limmathurotsakul *et al*, 2010). The causative agent of melioidosis is a Gram-negative bacterial pathogen, *Burkholderia pseudomallei* (White, 2003).

In addition, an anthrax attack in the United States in 2001, caused by contamination by bacterial spores of *Bacillus anthracis*, has raised public awareness of health security, and in particular the threat of bioterrorism (Pfeiffer, 2009). Due to the high mortality of *B. pseudomallei* infections and its persistence in soil and ground water, the bacteria has the potential to be used as a bioweapon. Harmful and serious disease causing microorganisms have been categorised by the Centers for Disease Control and Prevention (CDC) and *B. pseudomallei* is listed as a category B bioterrorism agent (Rotz *et al*, 2002). Indeed, the closely related bacteria, *B. mallei*, which causes the disease glanders, primarily found in horses and also affecting other animals and humans, was successfully used as a biological weapon during World War I by infecting livestock for export to Allied forces (Hawley & Eitzen, 2001).

1.1. Microbiological aspect of *B. pseudomallei*

B. pseudomallei is a non spore forming, bacillus-like, motile, aerobic bacterium. It is also a soil-dwelling saprophyte and frequently found in drenched rice paddy fields in endemic regions of melioidosis (White, 2003). Unlike other *Burkholderia* species, it is oxidase positive and capable of arabinose assimilation (Cheng & Currie, 2005). Ashdown's selective agar (Ashdown, 1979) is typically used in order to isolate and grow *B. pseudomallei* from clinical specimens (Cheng *et al*, 2006). Various types of colony morphology of *B. pseudomallei* have been observed and the association of morphotype and phenotype changes was demonstrated by Chantratita *et al* (2007a). Seven unique colony morphotypes have been identified according to its surface texture of the colony center, outer edge of colony, surface roughness in outer half of colony, colony diameter and colony color (Figure 1.1).

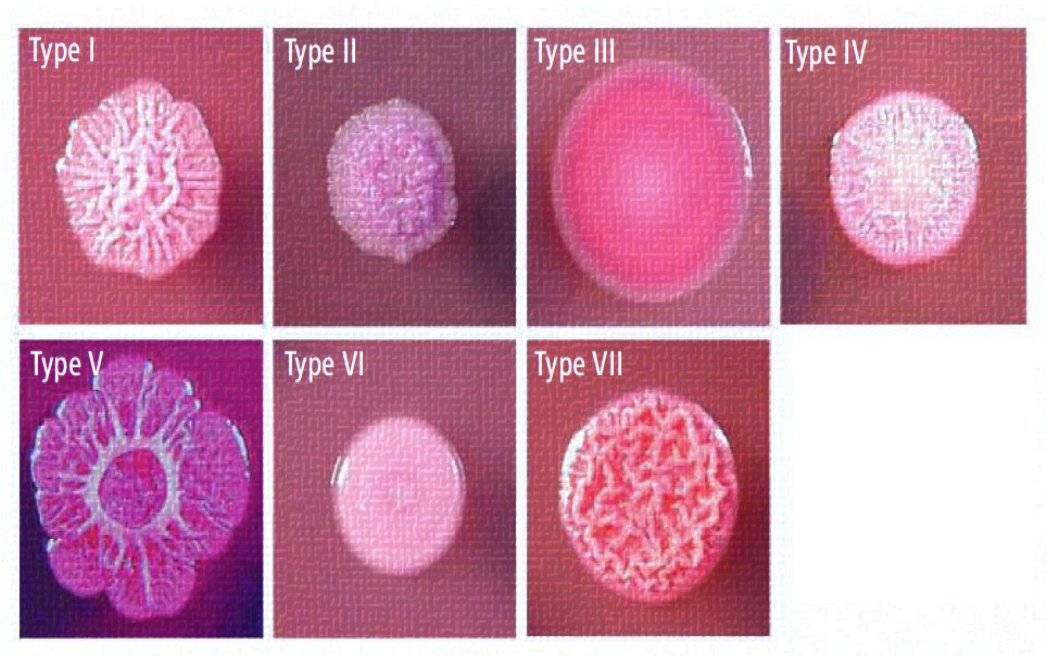


Figure 1.1 Photographs showing the variation of colony morphotypes of *B. pseudomallei* grown on Ashdown's selective media. Approximately 75% of clinical samples from melioidodsis patients are found as type I colonies, which has a rough texture at the center and is irregular in the outer edge, with a pale purple colour. This figure is adapted from Stone (2007).

B. pseudomallei has the ability to survive in different harsh environmental conditions including low moisture content soil, acidic environments, the presence of detergents (Cheng & Currie, 2005) and even a lack of nutrients (Wuthiekanun *et al*, 1995). It can also survive in distilled water for at least 16 years (Pumpuang *et al*, 2011). A study of *B. pseudomallei* survival in water showed that a putative membrane protein, BPSL0721, which was up regulated, may be involved in the adaptation for surviving in an aqueous environment. However, this gene is not essential for survival as the *BPSL0721* deficient mutant was still able to survive in water (Moore *et al*, 2008). The effect of chemical and physical environments on *B. pseudomallei* survival and persistence has been reviewed by Inglis & Sagripanti (2006). *B. pseudomallei* can survive in various conditions of high water content, pH changes, osmotic pressure and chemical stress, but is less tolerant to ultra-violet light exposure. Another factor that influences disease outbreaks is the weather conditions, with many cases of melioidosis occurring in the rainy season (Wiersinga *et al*, 2006). As *B. pseudomallei* is susceptible to ultraviolet light (Sagripanti *et al*, 2009), this may explain the relatively low incidence of melioidosis during the dry season when natural sunlight levels are very high.

1.2. Geographical distribution of melioidosis

The traditionally recognised endemic areas of melioidosis are approximately the tropical latitudes between 20°N and 20°S (Cheng & Currie, 2005). The hyper-endemic region includes northeast Thailand, northern Australia, Singapore and parts of Malaysia (Figure 1.2). Recently, the endemic zone has been extended to cover Papua New Guinea, the Indian subcontinent, southern China, Hong Kong and Taiwan. In addition, sporadic cases have been reported from both the Americas and Africa (Currie *et al*, 2008). International travel makes *B. pseudomallei* infection and melioidosis an increasing global concern.

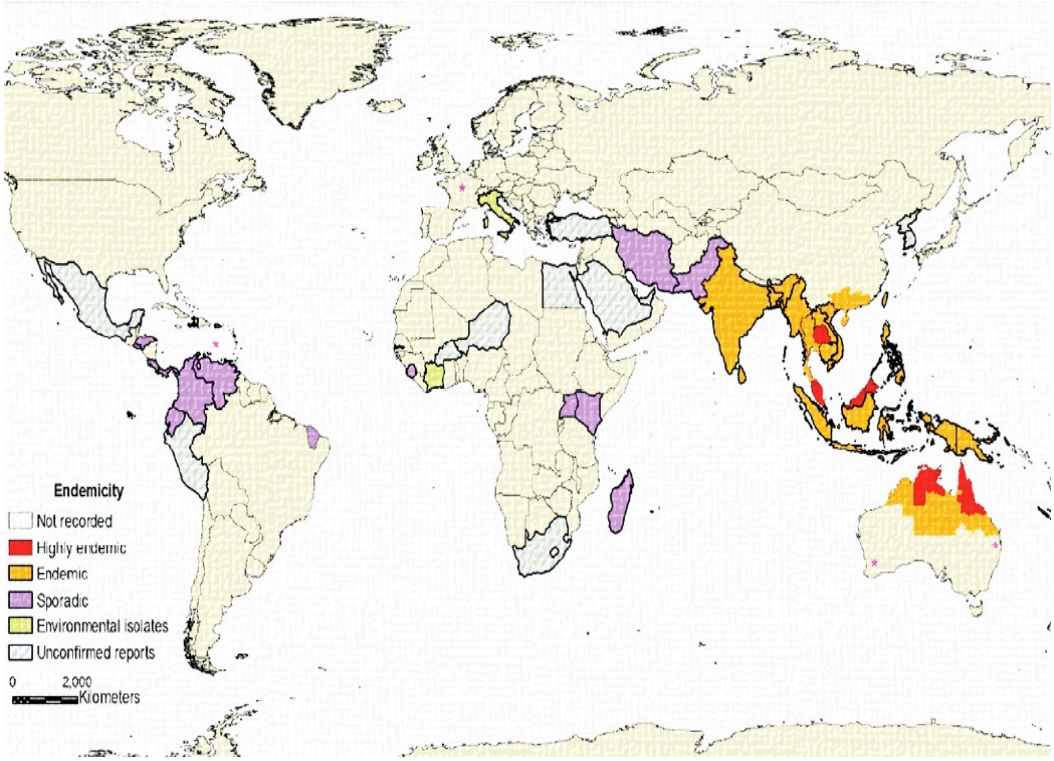


Figure 1.2 The worldwide distribution map of *B. pseudomallei* and melioidosis. Major zones of endemic disease are colored in red, whereas yellow areas represent the spread of melioidosis. This figure is taken from Currie *et al* (2008).

1.3. Clinical presentation

Melioidosis can be acquired by percutaneous inoculation, inhalation or ingestion, with inoculation thought to be the major route for disease infection (Wiersinga *et al*, 2006). The latency period can last from a few days to up to 62 years and clinical features may vary from an asymptomatic state, to acute or chronic lung infection, or to overwhelming septicemia (Woods *et al*, 1999). Apart from the lung, which is the most commonly affected organ, skin infections, bone and join infections, liver and spleen abscesses, genio-urinary infection and brain abscess can be found as clinical symptoms of melioidosis (Figure 1.3). A variation in the clinical pattern of melioidosis in Australia, Thailand, Singapore and Malaysia has been observed (Cheng & Currie, 2005). Misdiagnosis of this disease frequently occurs due to its polymorphic symptoms, which are similar to other lung infections, and *B. pseudomallei* is recognised as “the great mimicker” (Stone, 2007).

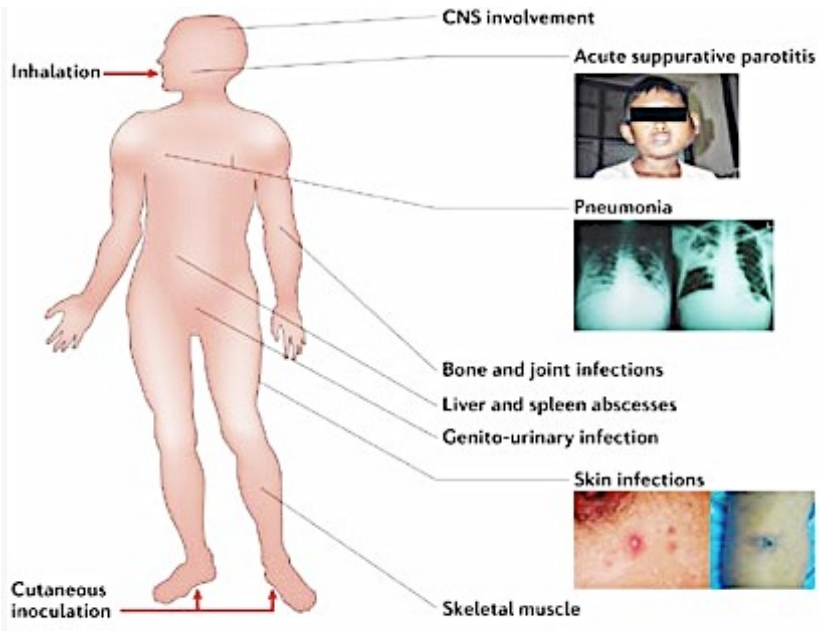


Figure 1.3 An illustration of clinical presentations of melioidosis: Melioidosis sepsis is associated with the bacterial dissemination to principle organs such as the lungs, liver and spleen, as well as the central nervous system (CNS). This figure is taken from Wiersinga *et al* (2006).

If the primary infection is survived, a major cause of morbidity and mortality is recurring melioidosis. The re-occurrence is about 6% in the first year and 13% throughout 10 years of medical follow-up (Peacock, 2006). Interestingly, 25% of recurrent cases were re-infection by different strains, which could imply differing levels of response to antibiotics within the strains (Maharjan *et al*, 2005). A study of clinical risk factors showed that 12-16 weeks of antibiotic treatment in melioidosis patients decreased risk of relapse, compared to those who were treated for 8 weeks or less (Limmathurotsakul *et al*, 2006). Different proteins expressed from *B. pseudomallei* isolated during primary and relapsing melioidosis have been identified using proteomic analyses (Velapatino *et al*, 2012). Up-regulation of *B. pseudomallei* proteins, including the HSP20/alpha crystalline family protein (BPSS2288) and flagellin (BPSL3319) in the relapsing isolate, compared to the initial isolate has been observed. These proteins could be determinants of melioidosis relapse and may be involved in melioidosis latency and virulence.

In addition, immunosuppressive people are at risk to the disease, especially those suffering from diabetes mellitus (Suputtamongkol *et al*, 1999). Between 37% and 60% of melioidosis patients are also diabetic. Exposure to soil or water by rice farmers in Thailand, and Aborigines in Australia, is also associated with contracting melioidosis. Other clinical risk factors for developing melioidosis such as excess alcohol consumption, chronic lung disease and renal disease have also been reported (Cheng & Currie, 2005), as well as heart disease, malignancy and age ≥ 50 years (Currie *et al*, 2010).

1.4. *B. pseudomallei* genome

To date, the completely sequenced genomes of four strains of *B. pseudomallei* are available in published databases, while another 19 have been drafted (Holden, 2009). The genome of *B. pseudomallei* strain K96243 (Figure 1.4) consists of two chromosomes of 4.07 megabase pairs and 3.17 megabase pairs (Holden *et al*, 2004). Many core genes involved in essential cellular functions, for example cell growth and metabolism, are found in the large chromosome whereas accessory genes associated with adaptation and survival in various environmental niches are found in the smaller chromosome.

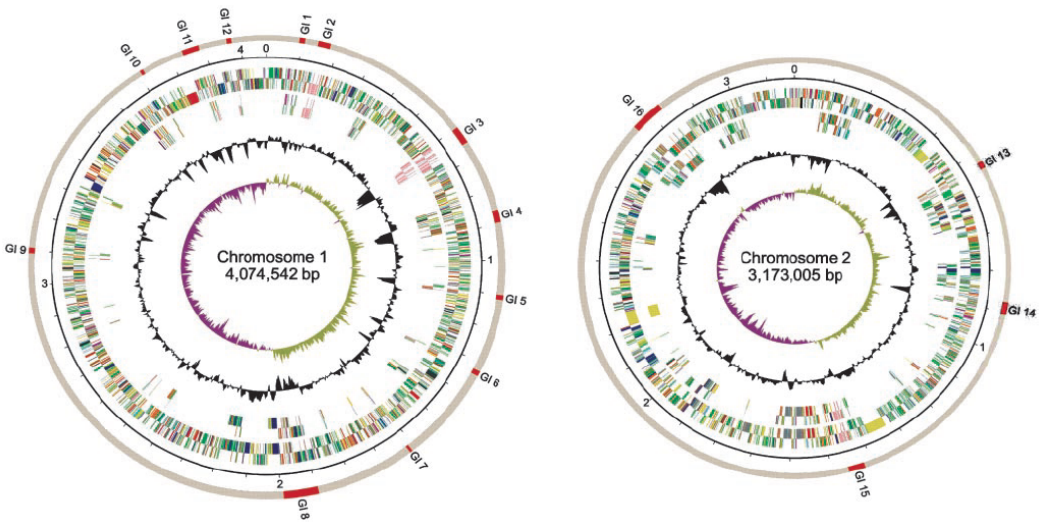


Figure 1.4 Schematic diagrams of two chromosomes in the *B. pseudomallei* strain K96243 genome: 3,460 coding sequences are encoded by chromosome 1 whereas 2,395 coding sequences are encoded by chromosome 2. This figure is taken from Holden *et al* (2004).

Interestingly, only 86% of the genome of *B. pseudomallei* strain K96243 (4619 genes) have been found in all strains, with 14% of the genome varying across different strains (Sim *et al*, 2008). This plasticity of the *B. pseudomallei* genome has been studied and genomic islands related to horizontal gene transfer have been identified as one of the crucial source of genomic diversity within *B. pseudomallei* (Tuanyok *et al*, 2008). In addition, the genomic variation of different *B. pseudomallei* isolates and gene loss in genomic islands of *B. pseudomallei* genomes have been observed (Lye *et al*, 2010). However, from a comparative study between the clinical and environmental isolates of *B. pseudomallei*, there is no evidence to confirm the association of genomic islands with the mechanism of virulence (Bartpho *et al*, 2010).

Since the completely sequenced first genome of *B. pseudomallei* strain K96243 was published (Holden *et al*, 2004), there has been a considerable emphasis on molecular based research to make insights into the pathogenicity and virulence mechanisms of *B. pseudomallei*. However, the ambiguous system of *B. pseudomallei* infection is still a key challenging research question.

1.5. Pathogenesis and virulence

1.5.1. Pathogenesis

A key element of the pathogenesis of *B. pseudomallei* is its ability to invade, survive and replicate within both phagocytic and non-phagocytic cells of the host (Allwood *et al*, 2011). Adhesion to the external surface of host cells is considered the first step in the pathogenesis of *B. pseudomallei* prior to an intracellular invasion. The initial attachment of *B. pseudomallei* to host cells is mediated by a bacterial capsule (Galyov *et al*, 2010). Subsequently, an intracellular invasion of *B. pseudomallei* occurs either in macrophages or epithelial cell lines. It is likely that the type III secretion system (T3SS) plays an important role in the bacterial invasion as well as escaping from endocytic vesicles (Stevens & Galyov, 2004). Once inside the host cell cytoplasm, the ability to survive intracellularly is a key for developing the infection while avoiding host immune surveillance. Strategies of evasion of the host immune response by *B. pseudomallei* include escaping from the endosome to the cytosol, inducing macrophage cell death and cell-to-cell

spreading through actin polymerization (Gan, 2005). *B. pseudomallei* can form actin based membrane protrusions by continuous nucleation of actin at one pole of the bacterial cell, allowing intracellular motility (Figure 1.5). The bacterial movement to adjacent cells leads to cell fusion and the formation of multinuclear giant cells (MNGC), which indicates the progression of infection (Adler *et al*, 2009).

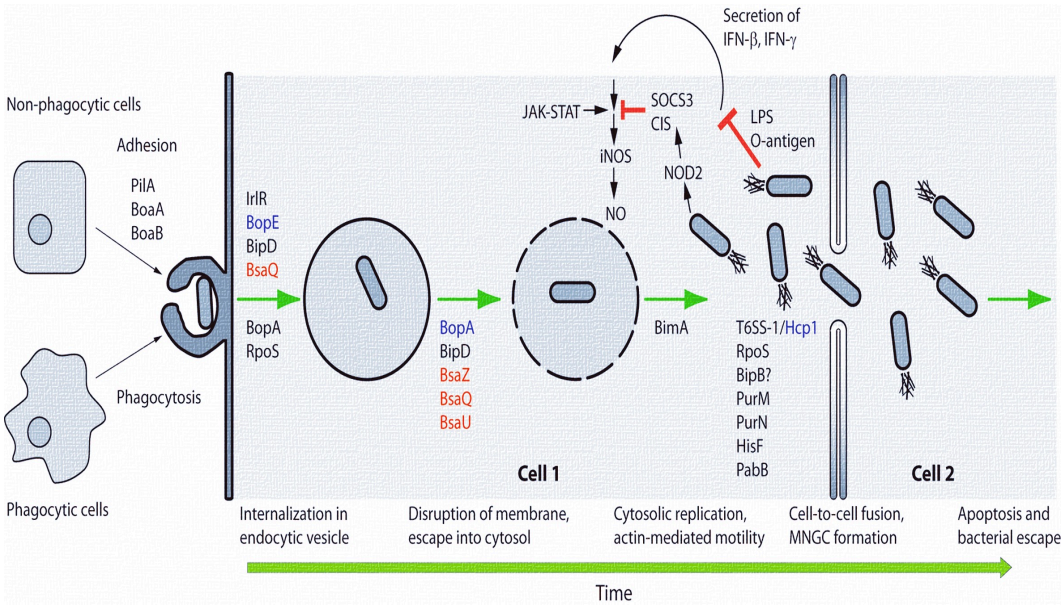


Figure 1.5 A schematic diagram of the *B. pseudomallei* intracellular lifestyle: An initial attachment to host cells involves the bacterial capsular polysaccharide, pili and adhesins. Molecular syringes like T3SS facilitate bacterial invasion. *B. pseudomallei* possesses intracellular survival mechanisms against host immune responses. Actin based mobility of the pathogen allows cell-to-cell spread and MNGC formation. This figure is adapted from Allwood *et al* (2011).

1.5.2. Virulence determinants

Some putative virulence factors of *B. pseudomallei* have been identified on the basis of existing evidence from a known virulent role in other pathogens. Quorum sensing, a communication system that depends on cell density in Gram-negative bacteria using N-acyl-homoserine lactones (AHLs) for coordinating gene

expression, is one such potential putative virulence factor. It plays a role in regulating a multidrug efflux pump responsible for antibiotic resistance in *B. pseudomallei* (Popat *et al*, 2008). Another recognised virulence determinant is the type III secretion system (T3SS). Proteins encoded from some T3SS gene clusters in *B. pseudomallei* are similar to those produced from *Salmonella typhimurium* and *Shigella flexneri* that are involved in pathogenesis (Troisfontaines & Cornelis, 2005). In addition, *B. pseudomallei* can produce an extracellular capsular polysaccharide that is related to virulence as well as lipopolysaccharide**,** flagella**,** pili and other secreted proteins including haemolysin lipases and proteases (Wiersinga *et al*, 2006). Recently, the type VI secretion system (T6SS) has been considered important in the intracellular lifestyle in *B. pseudomallei* (Galyov *et al*, 2010).

1.5.2.1. Capsular polysaccharide

A homopolymer of -3)-2-O-acetyl-6-deoxy-β-D-manno-heptopyranose(1- forming an extracellular capsular polysaccharide in *B. pseudomallei* has been identified as a possible virulence determinant (Reckseidler *et al*, 2001). The capsular polysaccharide may block the access of complement receptor-1 on phagocytes to C3b deposited on the bacterial surface, resulting in host immune response failure and thus *B. pseudomallei* survival in serum (Reckseidler-Zenteno *et al*, 2005). Capsule mutants of *B. pseudomallei,* prepared by the deletion of *wcb*, a capsular polysaccharide encoded gene operon, have revealed an attenuation in mice (Warawa *et al*, 2009).

1.5.2.2. Lipopolysaccharide

A comparative study of the lipopolysaccharide profiles of *B. pseudomallei* (pathogenic) and *B. thailandensis* (non-pathogenic) suggests that lipopolysaccharide may not be involved in virulence in *B. pseudomallei* (Wiersinga *et al*, 2006). However, a knockout of one of the genes encoding lipopolysaccharide in *B. pseudomallei* exhibits an attenuation in hamsters, and serum killing (DeShazer *et al*, 1998).

1.5.2.3. Quorum sensing

An intrinsic bacterial resistance to antibiotics is associated with efflux pumps such as AMrAB-OprA, BpeAB-OprB and BpEF-OprC, which secrete the signaling molecules, N-acyl-homoserine lactones (AHLs). In *B. pseudomallei*, there are up to six different types of AHLs, the composition of which may vary slightly from strain to strain (Chan *et al*, 2007). The presence of AHLs in *B. pseudomallei* cultures has been reported with an involvement of the LuxI protein and LuxR transcriptional regulators in AHL biosynthesis (Wiersinga *et al*, 2006). *LuxIR* quorum sensing homologue mutants of *B. pseudomallei* show a decreasing virulence in hamsters and mice, compared to the wild type strain (Ulrich *et al*, 2004).

1.5.2.4. Type III secretion system (T3SS)

The T3SS is a bacterial machinery that allows direct communication between bacteria and host cells and injects effector proteins into the host cytosol across the cellular membrane in the presence of translocator proteins. The T3SS of *B. pseudomallei* belongs to the Inv-Mxi-Spa family, which is known as a key virulence determinant in *S. enterica* and *S. flexneri* (Troisfontaines & Cornelis, 2005). In *B. pseudomallei*, T3SS is encoded by the *Burkholderia* secretion apparatus (*bsa*) gene cluster. Compared to the wild type strain, the deficient T3SS mutants exhibit a reduction of replication in J774.2 macrophage cells, a lack of ability to escape from endocytic vacuoles and an absence of membrane protrusion and actin tail formation (Stevens *et al*, 2002).

1.5.2.5. Type VI secretion system (T6SS)

In the *B. pseudomallei* genome, fourteen T6SS orthologs have been found and they are similar to those present in *B. mallei* (Shrivastava & Mande, 2008). Six gene clusters of T6SS are encoded in the *B. pseudomallei* K96243 genome. They have been assigned as T6SS-1 (BPSS1496 to BPSS1511), T6SS-2 (BPSS0515 to BPSS0533), T6SS-3 (BPSS2090 to BPSS2109), T6SS-4 (BPSS0166 to BPSS0185), T6SS-5 (BPSS0091 to BPSS0117), and T6SS-6 (BPSL3096 to BPSL3111) according to the T6SS of *B. mallei* described by Schell *et al* (2007).

The T6SS-1 cluster genes were inducible following *B. pseudomallei* invasion of macrophages, but they were not involved in bacterial survival and growth inside macrophages (Shalom *et al*, 2007). A deficiency of multinucleated giant cell (MNGC) formation was observed in *B. pseudomallei* T6SS-1 mutants. In addition, Δhcp1 mutants decreased *B. pseudomallei* intracellular growth and cytotoxicity in RAW 264.7 macrophages (Burtnick *et al*, 2011). Furthermore, in a Syrian hamster model of virulence, the T6SS-1 was shown to be a crucial determinant of virulence, based on the LD_{50} against the relevant mutants of *B. pseudomallei*, whereas the other five T6SS gene clusters did not. These results suggest that the T6SS-1 is involved in virulence and plays a significant role in the intracellular lifestyle of *B. pseudomallei*.

1.5.2.6. Flagella

B. pseudomallei is motile and flagellated. The difference between the flagellin gene *fliC* in *B. pseudomallei* and *B. thailandensis*, a closely relative non-pathogenic strain may be related to virulence. A *B. pseudomallei flic* knockout mutant exhibits less invasive abilities into a mouse macrophage cell line, RAW 264.7, and a human lung epithelial cell line, A549, compared to the wild type strain, but higher than that of *B. thailandensis* (Chuaygud *et al*, 2008). These results suggest that the flagella in *B. pseudomallei* facilitate bacterial invasion in both phagocyte and non-phagocyte cells. Based on the LD_{50}, the flagella defective mutant, which was non-motile and aflagellate, was less virulent during intranasal infection of BALB/c mice comparing to the wild type *B. pseudomallei* (Chua *et al*, 2003), which suggests that flagella are a virulence determinant of *B. pseudomallei* infection in this mouse model.

1.5.2.7. Pili

Type IV pili associated with adherence in *B. pseudomallei* has been suggested to play a role in virulence. A deletion mutant of the gene encoding the putative pilus structural protein (PilA) has exhibited a reduction of adherence to human epithelial cells, resulting in a delay of death in nematodes. Infection by this mutated strain was also attenuated in BALB/c mice following intranasal infection in comparison with the wild type *B. pseudomallei* (Essex-Lopresti *et al*, 2005).

1.6. Host-pathogen interaction

1.6.1. Pathogen recognition

When a pathogen strikes, the host's innate immune system is a surveillance system which recognises microorganisms via a limited set of germline-encoded pattern-recognition receptors (PRRs) (Akira *et al*, 2006). PRRs have an ability to recognise microbial components, called pathogen-associated molecular patterns (PAMPs), which are essential for their survival and are difficult to alter. Different PRRs interact with specific PAMPs, regardless of the pathogen life-cycle stage. A family of commonly known PRRs is the toll-like receptors (TLRs) which are type I integral membrane glycoproteins. Extracellular domains of TLRs contain varying numbers of leucine-rich-repeat (LRR) motifs. The cytoplasmic domain of TLRs is known as the Toll-IL-1R domain, which is homologous to the cytoplasmic region of the interleukin 1 receptor (IL-1R) (Cook *et al*, 2004). TLRs are classified into several subfamilies based on their primary sequences. Each of these subfamilies recognises related PAMPs. In *B. pseudomallei* infection, PAMPs that are present on the bacterial surface interact with the corresponding TLRs of the host cell, for example, CD14-TLR4, peptidoglycans-TLR2, flagellin-TLR5 and bacterial DNA or CpG-TLR9 (Wiersinga *et al*, 2006).

1.6.2. Bacterial intracellular survival

The initial host defense against *B. pseudomallei* includes natural antibiotic peptides, the complement system and phagocyte cells. Antimicrobial peptides called defensins are naturally produced to kill bacteria attacking host cells. It is unfortunate that *B. pseudomallei* is resistant to the natural immunity (Jones *et al*, 1996). In addition, *B. pseudomallei* can produce an extracellular polysaccharide capsule that prevents complement attack. This polysaccharide capsule reduces the deposition of complement factor C3b, resulting in anti-phagocytosis (Reckseidler-Zenteno *et al*, 2005).

Macrophages are well-known phagocyte cells, whose functions are to ingest and destroy bacteria. Responses of macrophages to *B. pseudomallei* are different from other bacteria. An inducible nitric oxide synthase (iNOS), which directs the clearance of bacterial pathogens, is suppressed when the macrophages are infected with *B. pseudomallei*. This can explain why this bacteria can survive and multiply

inside host macrophage cells. In addition, IFN-γ, that has an important role in the host immune system, is produced at a lower level in the host's resistance to *B. pseudomallei* infection, compared to other bacterial infections, resulting in increased bacterial loads in melioidosis patients (Panomket, 2011).

1.7. Melioidosis management

The difficulties of melioidosis management include a limitation in diagnostic approaches and the increasing resistance of the bacteria to the current anti-infective treatments.

1.7.1. Diagnosis

Although a number of techniques have been developed for the rapid detection of *B. pseudomallei* in patient fluids, bacterial culture using selective media (e.g. Ashdown's agar and *B. pseudomallei* selective agar, BPSA) are still the gold standard for melioidosis case confirmation (Peacock, 2006). Although this traditional diagnostic method is a time-consuming process, it has the advantage of identifying high risk of death patients in terms of a quantitative culture (Wuthiekanun *et al*, 2010). Alternatives to this culture-based method, include serological diagnosis using antibody and antigen detection. An indirect haemagglutination assay (IHA) is commonly performed, although it has a poor sensitivity and specificity. This particular diagnostic method is problematic in Thailand, where there is a high rate of background seropositive population (Cheng & Currie, 2005). As an improvement to the serological tests for melioidosis diagnosis, it has been reported that a crude *B. pseudomallei* antigen enzyme-linked immunosorbent assay (ELISA) had an advantage over IHA observations (Chantratita *et al*, 2007b; Limmathurotsakul *et al*, 2011).

During 20 years of observation of melioidosis in Northern Australia, the mortality of melioidosis has decreased from 30% at the beginning to 9% in the last five years, which is attributed to earlier diagnosis and enhancements in intensive care management (Currie *et al*, 2010).

1.7.2. Antibiotic resistance

Burkholderia species are known to be one of the most antibiotic resistant bacteria. Predominantly, drug resistance mechanisms of bacteria involve enzymatic inactivation of antibiotics, drug target alteration and limited drug permeability into the bacterial cell (Poole, 2002). The latter plays a key role in antibiotic resistance mode of action in *Burkholderia.*

There are three mechanisms of action that limit antibiotic penetration into bacterial cells. Firstly, lipopolysaccharide (LPS), a unique component of Gram-negative bacterial cell walls, restricts the binding of cationic antibiotics, including aminoglycosides and polymyxins. Also, it renders the bacteria resistant to cationic peptides of the human innate immune system. Secondly, low cell membrane porin protein content may affect the permeability of water-soluble antibiotics into the bacterial cell due to an abnormal function. Thirdly, transport proteins, namely efflux pumps, are involved in antibiotic resistance. Antibiotics can be removed from the bacterial cell by the specific efflux pumps before being delivered to their cellular targets (LiPuma, 2007). In addition, biofilm formation may cause the bacteria to resist an antimicrobial agent. Possible resistant mechanisms in the biofilm include reduced drug diffusion through the biofilm, drug inactivation within the biofilm, the presence of less susceptible stationary-phase organisms, and the up-regulation of biofilm associated antibiotic resistance genes (LiPuma, 2007).

Commercially available antibiotics can be categorised into several classes, depending on their modes of action. β-lactams, including penicillins, carbapenem and cephalosporins are related to an inhibition of bacterial cell wall synthesis. Aminoglycosides, macrolides, tetracyclines and chloramphenicol are involved in the binding of the ribosome subunit, resulting in the inhibition of protein synthesis. Fluoroquinolones disrupt DNA replication as a result of DNA gyrase inhibition (Kohanski *et al*, 2010). On the whole, *B. pseudomallei* is intrinsically resistant to many antibiotics. A susceptibility of *B. pseudomallei* to 35 antimicrobial agents in terms of the minimum inhibitory concentration (MIC),

determined by agar dilution in Mueller–Hinton medium, has been studied by Thibault *et al* (2004). The results indicate that *B. pseudomallei* is extremely resistant to penicillins (amoxicillin, ticarcillin), aminoglycosides (gentamicin, tobramycin, netilmicin, amikacin), macrolides (erythromycin), cephalosporins (cefoxitin), fluoroquinone (norfloxacin), fosfomycin, and clindamycin. The mechanism of resistance to aminoglycosides and macrolides can be explained by the presence of a multidrug efflux system in *B. pseudomallei* (Van Bambeke *et al*, 2000). The BpeAB-OprB pump in *B. pseudomallei* was accountable for the efflux of the aminoglycosides gentamicin and streptomycin, the macrolide erythromycin and the dye acriflavine (Chan *et al*, 2004).

In addition, further antibiotics that *B. pseudomallei* is moderately resistant to include fluoroquinones (nalidixic acid, pefloxacin, ciprofloxacin, gatifloxacin and levofloxacin), co-trimoxazole and rifampicin. These antimicrobial agents are involved in the disruption of DNA replication and RNA synthesis. On the other hand, carbapenems like imipenem, the third generation of cepharosporins (ceftazidime and cefotaxime), other β-lactams (piperacillin and piperacillin/tazobactam and co-amoxiclav) tetracyclines (minoxycycline and doxycycline), and chloramphenicol, are efficient for killing this pathogen.

Recently, resistance to ceftazidime during antimicrobial therapy, which is currently used as the first line treatment for melioidosis, has been reported (Chantratita *et al*, 2010). These observations mean that finding new therapeutic drugs for developing melioidosis treatments are an urgent need.

1.8. Identification of potential therapeutic target proteins

In attempts to identify proteins involved in infection and virulence mechanisms of *B. pseudomallei*, that may be targeted by novel therapies, many studies have been carried out using different techniques. 312 putative essential genes of *B. pseudomallei* have been identified by a computational bioinformatics approach (Chong *et al*, 2006). *B. pseudomallei* protein sequences were compared with known essential genes encoding proteins from other bacterial pathogens, and also against human protein sequences to identify targets that did not have human homologues. The subset of genes thus identified have been annotated as vital genes for *B. pseudomallei* survival, and possible potential candidates for antimicrobial drug target development. In addition, the relative levels of protein expression in *B. pseudomallei* and the non-pathogenic *B. thailandensis* have been observed by proteomic analysis (Wongtrakoongate *et al*, 2007). Several proteins in *B. pseudomallei* that are up regulated compared to *B. thailandensis* have thus also been identified as potential markers of pathogenicity and virulence. They may play a key role in virulence as they were highly expressed in the virulent species but missing in the avirulent species.

Su *et al* (2008) have identified a set of *B. pseudomallei* target proteins that raise the human immune response in infected patients. The genomic expression library of a clinical strain of *B. pseudomallei* D286 was constructed and immunologically screened with pooled melioidosis infected human sera. 109 expressed polypeptides reacted with melioidosis positive sera and the coding sequence of sero-positive clones were analysed for sequence identification. These identified proteins are involved in cell envelope biogenesis, cell mobility and secretion, transcription, metabolism, transportation and also include many uncharacterised proteins. Many of these proteins (30%) are found in the cytoplasm. With an assumption that the cytoplasmic localised proteins are exposed after destruction of bacterial cells by the host immune defense, the results suggest that these identified immunogenic proteins may be induced and up-regulated during human infection. These immunogens could be further investigated for anti-infective targets, and a number of them have been investigated in this thesis.

1.9. Structural genomics

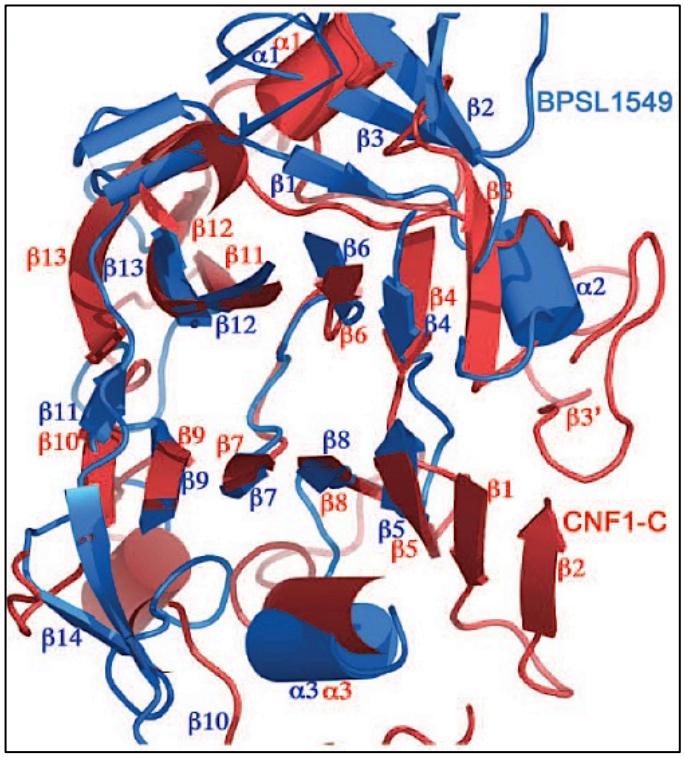
A fundamental of structural genomics of proteins with unknown function is to start from the gene sequence, produce protein and determine its three-dimensional structure, which could provide useful biological information of the protein (Thornton *et al*, 2000). Structural genomics requires techniques such as protein crystallography and nuclear magnetic resonance (NMR) to determine the structure of biologically interesting molecules from model organisms and those with medical importance, aiming to derive function from the structures (Brenner, 2001). From the determined structure, the organization of the protein chain in three dimensions is revealed and the residues which are buried in the core or involved solvent exposure on the protein surface can be identified as well as the overall shape and electrostatic properties of the protein. The protein structure can also reveal the quaternary structure in the crystal conditions or in solution at high concentration, which may indicate its biological multimeric state. In addition, biochemical and functional information can be extracted from the structure of protein-ligand complexes, that provide a crucial clue to their catalytic mechanisms. The similarity of the protein folds and structural motifs of an uncharacterised protein compared to the structural information of characterised proteins in the available databases may be used to infer its biochemical and biological functions.

In one such study, the structure determination of BPSL1549, an uncharacterised protein from *B. pseudomallei* has led to the discovery that it is a novel toxin (Cruz-Migoni *et al*, 2011). According to proteomic profiles, BPSL1549 has been identified as a putative virulence factor since the protein is expressed in *B. pseudomallei* but absent in *B. thailandensis*, a non-pathogenic strain (Wongtrakoongate *et al*, 2007). At the protein sequence level, BPSL1549 has no sequence similarity to any protein excluding other *Burkholderia* species. The crystal structure of BPSL1549 has been determined and analysed (Cruz-Migoni, 2011). By comparison to available structures deposited in the protein data bank, the fold recognition revealed that BPSL1549 is structurally similar to the catalytic domain of the cytotoxic necrotizing factor 1 (CNF1-C) from *Escherichia coli* (Figure 1.6a). And, indeed, the three crucial catalytic residues of CNF1-C are

present in BPSL1549 and are virtually the only conserved residues between these two proteins (Figure 1.6b). CNF1-C is an enzyme that catalyses the deamidation of a specific glutamine residue into a glutamic acid in the Rho protein family, which regulates the GTPase involved in actin cytoskeleton assembly. This comparative structure information suggested that BPSL1549 could also have a toxic role similar to CNF1-C. Further investigations have shown that BPSL1549 specifically deamidates glutamine-339 of the eukaryotic transcription factor eIF4A, rendering it inactive and thus inhibiting protein translation, ultimately resulting in cell death (Cruz-Migoni *et al*, 2011). These results support the proposal that BPSL1549 could play important roles in pathogenesis and virulence in *B. pseudomallei*. This study shows that structural analysis of a protein of unknown function can lead to important insights into its biological function and can shed light on pathogenicity and virulence mechanisms.

Structural genomics also has an impact on drug discovery. The availability of the three-dimensional structure of drug targets facilitates pharmaceutical development, especially for structure-based drug design (Dry *et al*, 2000). Structure data allow an identification of active sites or ligand binding sites and the optimization of lead compounds to reveal the selectivity and specificity preferences of druggable molecules at atomic resolution. DrugBank targets by structural genomics centers have been listed in Weigelt *et al* (2008). Together with a bioinformatics approach, structural information extracted from the determined protein-ligand complex structure can be inferred and contributes to the development of drug candidates (Congreve *et al*, 2005).

a)



b)

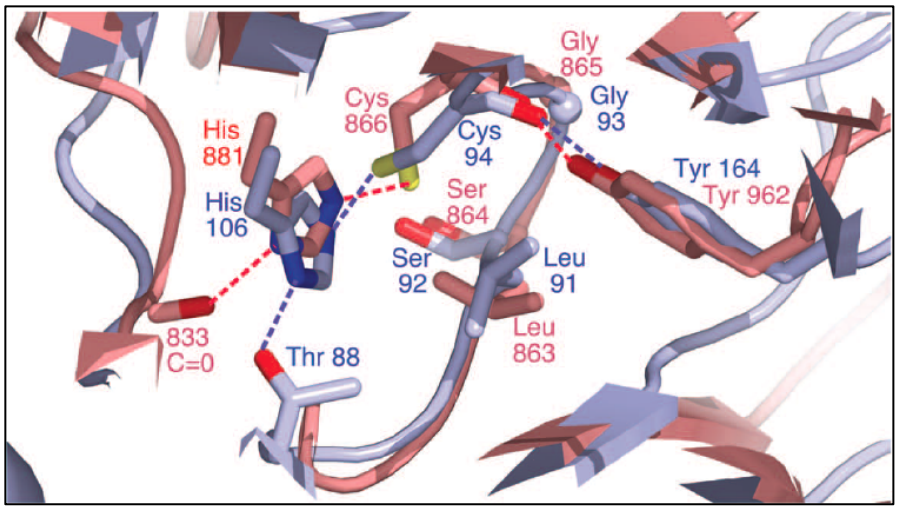


Figure 1.6 From structure to function: a) Superposition of BPSL1549 (blue) and CNF1-C (red) structures shows their structural similarity. b) Catalytic residues of CNF1-C are identical to equivalent residues in BPSL1549, leading to the biological function discovery of BPSL1549. This figure was taken from Cruz-Migoni *et al* (2011).

1.10. Aims of the study and thesis outline

This thesis describes the instigation of structural studies on several potential target proteins of *B. pseudomallei* for drug discovery and use of structural genomics, with an attempt to gain more understanding of the molecular mechanisms underlying the disease. The proteins that form the subject of this thesis are BPSL3022 (cell division protein FtsQ), BPSS0945 (subfamily M23 unassigned peptidase), BPSS0238 (putative penicillin binding protein) and the proteins of unknown function BPSS1416, BPSS0603, BPSS0683 and BPSL0606. The rationale behind this target selection is given in the introduction to Chapter 4.

The aims of this study are as followings.

- to create a list of potential target proteins of *B. pseudomallei*
- to clone the genes encoding these proteins
- to produce the proteins in *Escherichia coli*
- to purify the soluble proteins
- to crystallise the protein
- to determine the three dimensional structure of the proteins
- to analyse the structures obtained.

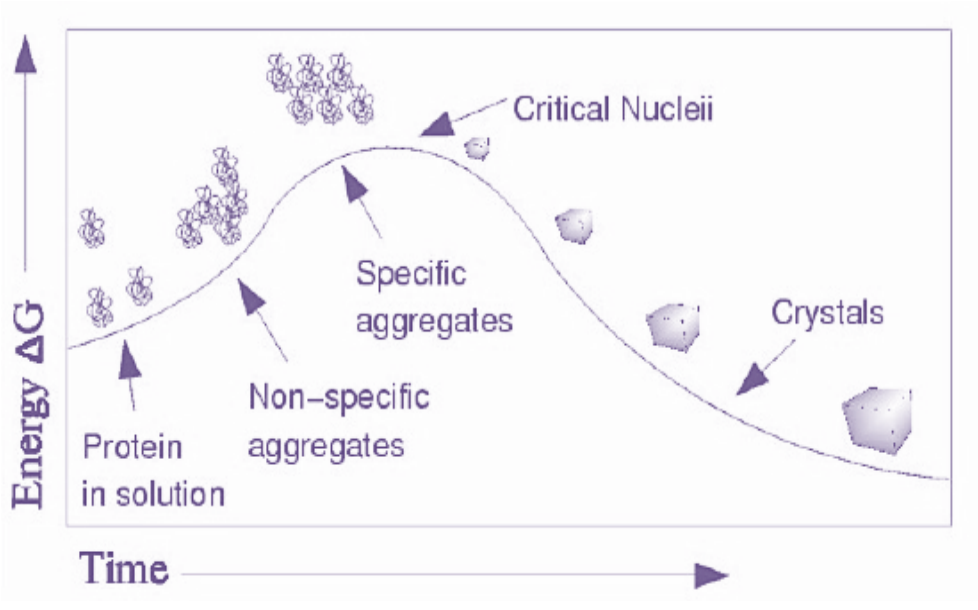
In this thesis, a brief description of protein crystallography theory involved in this study is summarised in Chapter 2. The materials and methods used through out this work are detailed in Chapter 3. Target selection, gene amplification and cloning, protein over expression and protein purification are described in Chapter 4. The crystallisation and structure determination of BPSL0606 is discussed in Chapter 5. Structure analysis of BPSL0606 is detailed in Chapter 6. Functional analysis of BPSL0606 is discussed in Chapter 7. Finally, conclusions and future works are included in Chapter 8.

Chapter 2: Protein Crystallography

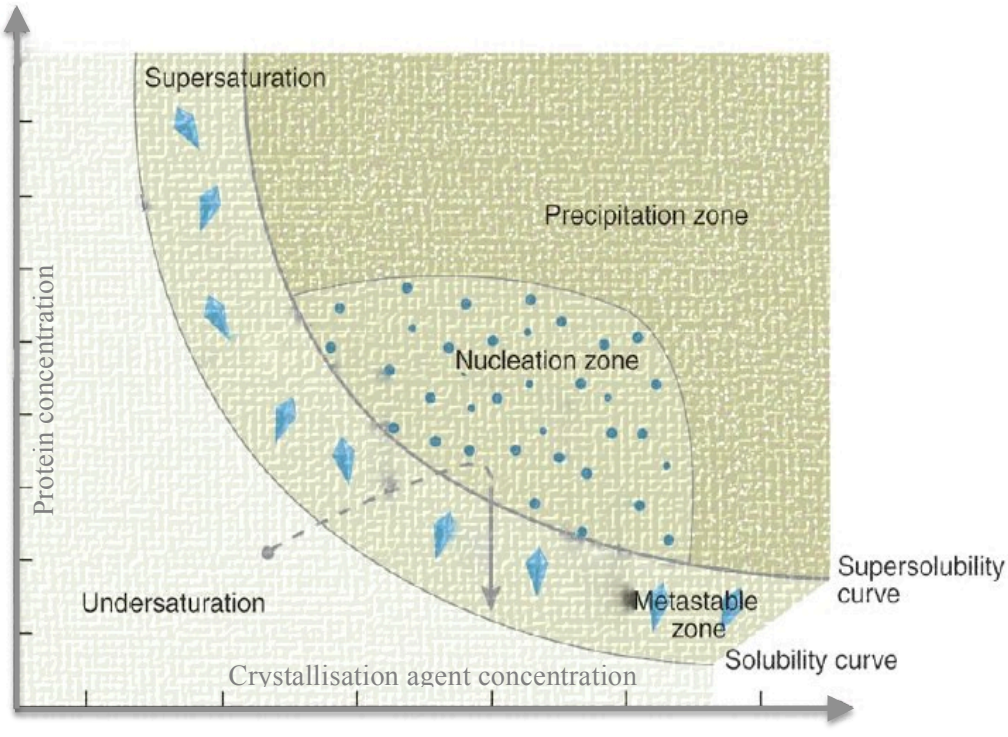
X-ray crystallography is a powerful technique for determining the three-dimensional structure of proteins and biological macromolecules. 88% of the entire biomolecular structures publicly deposited in the protein data bank (PDB) December 2012 release have been determined by X-ray crystallography. Although protein crystallography has advantages in terms of providing accurate models of protein structure at atomic resolution, with no limitation in the size of protein samples, it is dependent on obtaining protein crystals as a starting point. Crystallographic theory is covered in a number of textbooks (Blow, 2002; Rhodes, 2006; Rupp, 2010). This chapter will briefly introduce the processes in the experiments described in this thesis including crystallisation of the protein sample, crystal manipulation, X-ray data collection and processing, structure determination, structure refinement and validation of the structure.

2.1. Protein crystallisation

If a purified and homogeneous protein of interest can be produced, crystals of the protein can be obtained as the protein solution moves toward a metastable supersaturated state. The crystallisation trials are subjected to multi-dimensional conditions to induce the nucleation. According to the thermodynamics of crystallisation, molecular collisions in the supersaturated protein solution drive phase separation and the nucleation event occurs when the physicochemical reaction reaches the critical free energy and overcomes the kinetic barrier in which crystals form, and the system proceeds to binary phase equilibrium (Figure 2.1a). This process involves protein solubility, which is affected by several parameters including pH, ionic strength, temperature and concentration of salts, organic precipitants and additives. An initial optimization of these variables for an individual protein to be crystallised is accomplished by combinatorial screening. Commercially available screens facilitate the coverage of crystallisation space conditions. In addition, programmed robots are commonly set up for automated crystallisation, where the experiments are miniaturised to nano-scale volumes reducing the protein requirements.



a)



b)

Figure 2.1 Crystallisation diagrams: a) A free energy diagram for crystallisation showing that nucleation requires free energy from the physicochemical reaction to overcome the kinetic barrier in order to form crystals. The figure is taken from McCoy (2010) and b) A schematic illustration of a phase diagram of protein crystallisation using the vapour diffusion method. As a result of water vapour diffusion in the closed system, protein and precipitant concentrations increase toward the supersaturation condition, allowing nucleation and thus crystal formation. The protein concentration drops as the crystals sequester the protein. The figure is adapted from Chayen & Saridakis (2008).

2.1.1. Vapour diffusion techniques

Vapour diffusion is a conventional method for protein crystallisation. It involves mixing a small amount of the protein solution with the precipitant solution, usually in a 1:1 ratio. The technique provides a closed system allowing the protein mixture drop to equilibrate against a larger volume of reservoir solution that may contain buffer, salt and/or precipitant. The vapour diffusion of water and volatile compounds in the closed system occurs slowly. Water vapour from the protein mixture drop diffuses into the reservoir solution, resulting in a proportional increment of both the protein and precipitant concentrations in the drop. This process leads into a supersaturated state and enables crystallisation. If successful crystallisation is achieved, the protein concentration in the drop decreases as the crystals sequester the protein (Figure 2.1b).

2.1.1.1. Sitting drop vapour diffusion

A drop of protein solution mixed with crystallisation solution from the reservoir is placed on a microbridge over the reservoir, which is sealed with a greased rim and a cover slide (Figure 2.2a). This technique can be adapted to an automated crystallisation set up, with a robotic dispenser, allowing for miniaturization and using very small quantities of protein, approximately 200 nanolitres. Thus, sitting drop vapour diffusion is commonly used for the initial screening in 96-well plates.

2.1.1.2. Hanging drop vapour diffusion

Unlike the sitting drop technique, the protein mixture drop is placed on a siliconised cover slide that is inverted to face the reservoir, and again grease is used to seal the well (Figure 2.2b). This technique requires manual set up, and is often used for an optimization of crystal growth conditions. Drops of up to 10 microlitres can be easily used, and thus large crystals, which are easier to manipulate can sometimes be grown.

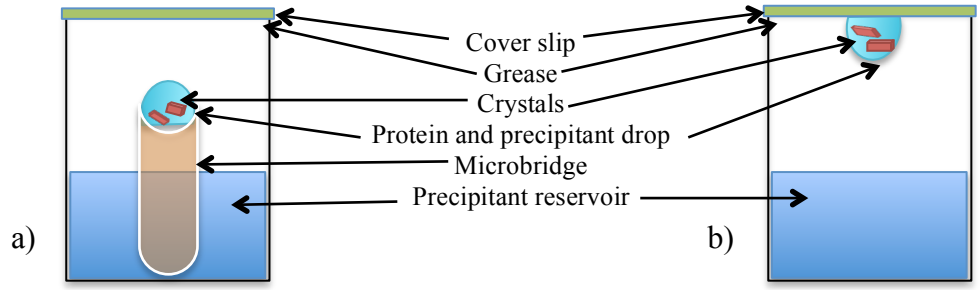


Figure 2.2 Illustrations showing the vapour diffusion experiments: a) Sitting drop technique- crystals are growing in the drop on the microbridge. b) Hanging drop technique- crystals are growing in the drop facing the reservoir.

2.1.2. Protein crystals

2.1.2.1. Crystal assembly

Protein crystals are assemblies of translatingly repeating identical units, called unit cells. These unit cells are parallelepipeds defined by lengths a, b, c and inter-axial angles α, β, γ. The unit cell is characterised by its space group, which defines the internal symmetry. There are 230 space groups, but as proteins are chiral, symmetry operators such as mirror planes cannot be used and there are thus only 65 possible space groups. The space group is defined by the lattice type and crystal system. There are five different lattice types, describing the position of the lattice points on the unit cell, where a lattice point is a point that is in exactly the same environment in all unit cells in the crystal. The primitive lattice (P) is a simplest type containing one lattice point at each vertex of the unit cell. An additional lattice point at the center of the unit cell generates the internal lattice (I). A centered lattice (C) has an additional point on one face and the face-centered lattice (F) has lattice points at the center of each face.

2.1.2.2. Crystal system

There are 7 crystal systems which are based on different rotational symmetries. Only 2-fold, 3-fold, 4-fold and 6-fold axes are allowed and each of these systems has different unit cell shape and dimensions. The simplest crystal system is triclinic with the unit cell $a \neq b \neq c$ and $\alpha \neq \beta \neq \gamma$. The cell in which $\alpha = \gamma = 90°$ and $\beta > 90°$ is called monoclinic, whereas the orthorhombic system has $\alpha = \beta = \gamma = 90°$. For unit cells with all three angles equal to 90°, if $a = b = c$, the system is

cubic and when a = b ≠ c, it is tetragonal. In addition, if the unit cell has a = b = c and α = β = γ ≠ 90°, it is rhombohedral, whereas the hexagonal unit cell has a = b = c and α = β = 90° γ = 120°. Together with centrosymmetry operators, these crystal systems create 14 Bravais lattices as shown in Figure 2.3. The combination of the lattice type, the crystal system and the presence of any screw axes (a combination of rotation and translation) gives the space group of the crystal.

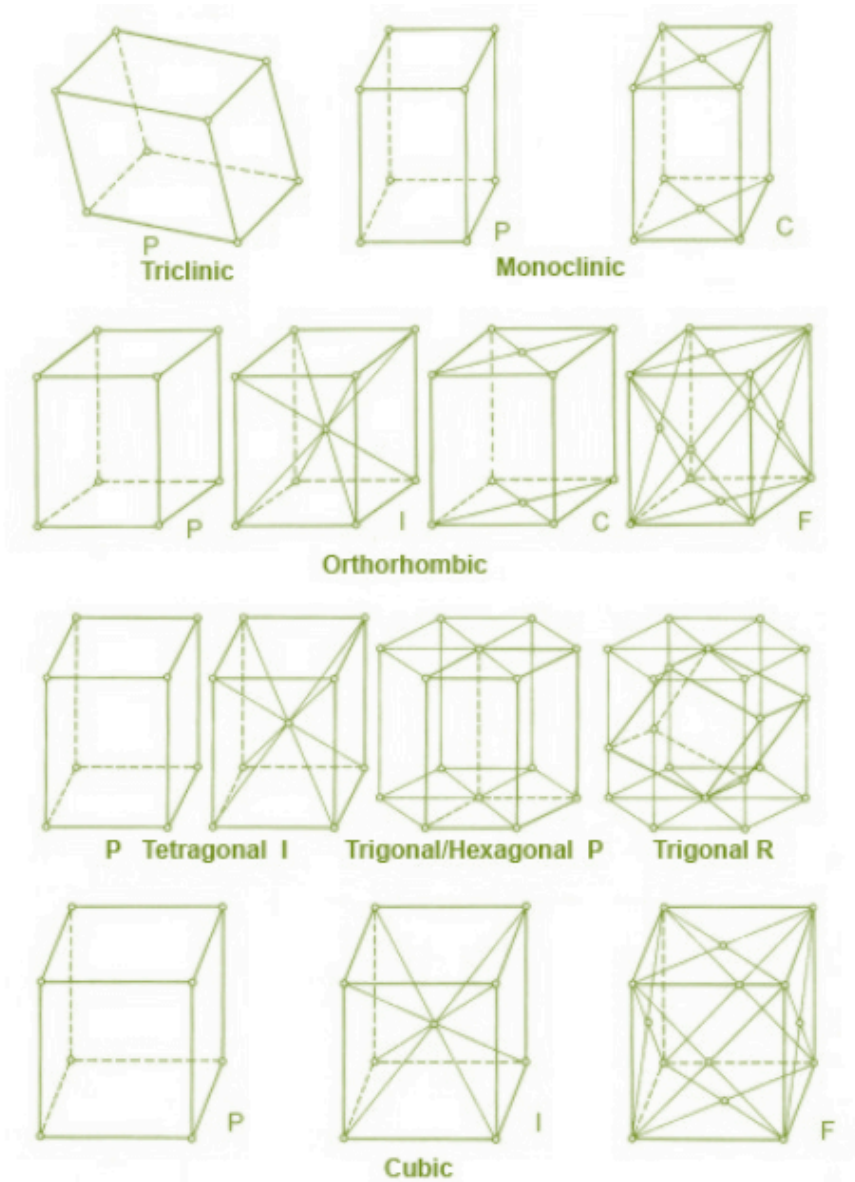


Figure 2.3 The 14 Bravais lattices: P = primitive, C = one face centered, I = body centered, F = face centered. This figure is taken from http://www.xtal.iqfr.csic.es/Cristalografia/index-en.html.

2.1.2.3. Mounting crystals

In order to harvest crystals from crystallisation trays, the protein crystals need careful manipulation. Because the crystal is in equilibrium with the surrounding mother liquor in the drop, it should be handled with the remaining mother liquor. Under a light microscope, the fragile crystal is gently fished out by a standard fiber loop with an appropriate size to the crystal and then transferred into a suitable cryobuffer, which is made from the solutions present in the mother liquor plus a cryoprotectant. The crystal is held in the loop by surface tension of the cryobuffer and ready to either plunged into liquid nitrogen or mounted onto the diffractometer. In this step, rapid transfer is preferred to avoid dehydration.

2.2. X-ray diffraction

X-ray diffraction data measurement from macromolecular crystals has lead to the determination of atomic structures. When a molecule is struck by an X-ray beam, the X-rays interact with the electrons in the molecule and diffraction occurs. However, this diffraction is very weak. A crystal of the molecule can be used to amplify the diffraction signal. The crystal contains many molecules in the same orientation and the diffracted rays from all the molecules have a cumulative effect, enabling the signal to be observed. The crystal diffracts the X-ray beam in many directions, resulting in a diffraction pattern. The electron density map derived from the diffraction pattern can be depicted by mathematical means of a three-dimensional periodic function.

2.2.1. Principles: Bragg's law

W.L. Bragg proposed an interpretation of X-ray diffraction as reflection on individual lattice planes hkl. In a crystal, a set of parallel planes, having Miller indices hkl, acts as a mirror resulting in a reflection. Each set of planes is an independent diffractor producing a single reflection. As Bragg's law defines (Figure 2.4), the condition to obtain the scattered X-ray in phase from parallel planes is

$$2d_{hkl} \sin \theta = n\lambda$$

where d_{hkl} is a lattice space between two parallel planes

θ is an angle of diffraction

λ is a wavelength of an incident X-ray

and n is an integer

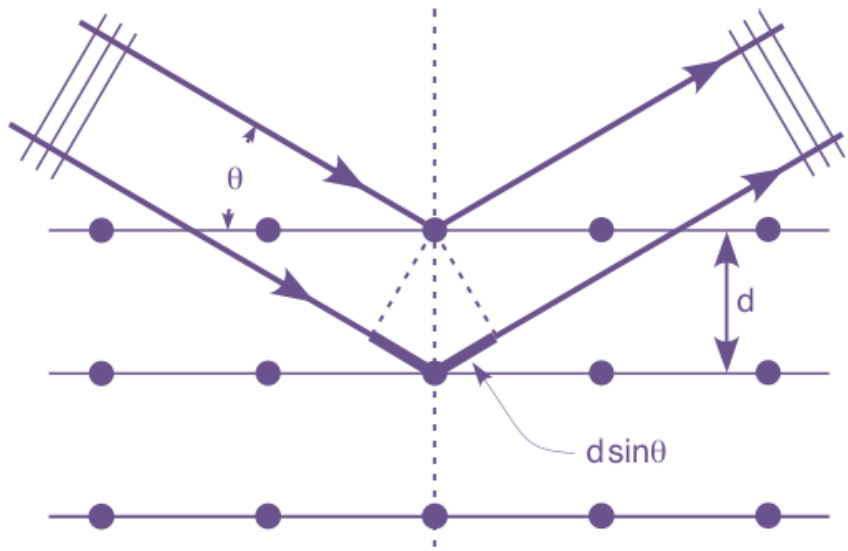


Figure 2.4 Bragg's law: A representation of the reflection of incident X-rays with an angle θ from a set of parallel planes (separated by a distance d) in the crystal to achieve the scattered X-ray in phase. The figure is taken from McCoy (2010).

Each spot in the diffraction pattern collected during the diffraction experiment represents a reflection related to the set of parallel planes with indices (hkl). h, k and l are defined as the numbers that the plane divides the unit cell edges (a, b and c), respectively.

2.2.2. Fourier Transform

A Fourier transform describes the relationship between molecules in the crystal and its diffraction pattern. A reflection can be depicted by a structure factor equation representing a diffracted X-ray. For a unit cell with n atoms, the structure factor for the reflection F_{hkl} can be expressed as a Fourier sum of all the atomic diffracting contributions:

$$\mathbf{F}_{hkl} = \sum_{j=1}^{n} f_j \, e^{2\pi i \, (hx_j + ky_j + lz_j)}$$

f_j is the scattering factor of atom j, dependent on the atomic number, and hence electrons, of the atom and is treated as a simple sphere of electron density. x_j, y_j and z_j represent the coordinates of atom j in real space, and h, k and l are the lattice indices of a specific reflection in reciprocal space. The contribution of individual atom j to $\mathbf{F}_{hkl}$ depends on both the atomic element, which determines the amplitude of the contribution (f_j), and the atomic position in the unit cell (x_j, y_j, z_j), which designates the phase of the contribution.

Alternatively, $\mathbf{F}_{hkl}$ can be expressed in terms of the volume element of electron density in the unit cell, ρ(x,y,z).

$$\mathbf{F}_{hkl} = \int_V \rho(x,y,z) \, e^{2\pi i (hx+ky+lz)} \, dV$$

Each volume element contributes to $\mathbf{F}_{hkl}$ derived from the integral over the unit cell volume, V denoted for all values of x, y and z. This equation shows that $\mathbf{F}_{hkl}$ is the Fourier transform of ρ(x,y,z) on the set of hkl planes. Therefore, all of the structure factor $\mathbf{F}_{hkl}$s constitute the transform of ρ(x,y,z) on all sets of parallel planes for the whole unit cell.

Due to the reversible nature of the Fourier transform, the electron density can also be described as the transformation of the structure factors as follows:

$$\rho(x,y,z) = \frac{1}{V} \sum_h \sum_k \sum_l \mathbf{F}_{hkl} \, e^{-2\pi i (hx+ky+lz)}$$

The calculation of electron density at a position (x,y,z) in the unit cell is performed by the summation over all the hkl planes. The structure factor $\mathbf{F}_{hkl}$ is a periodic wave function with amplitude, frequency and phase. It can also be represented by a complex vector and decomposed into its amplitude, $|\mathbf{F}_{hkl}|$ and phase, α_{hkl}. The electron density equation can then be expressed as follows:

$$\rho(x,y,z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}|\mathbf{F}_{hkl}|\ e^{i\alpha}e^{-2\pi i(hx+ky+lz)}$$

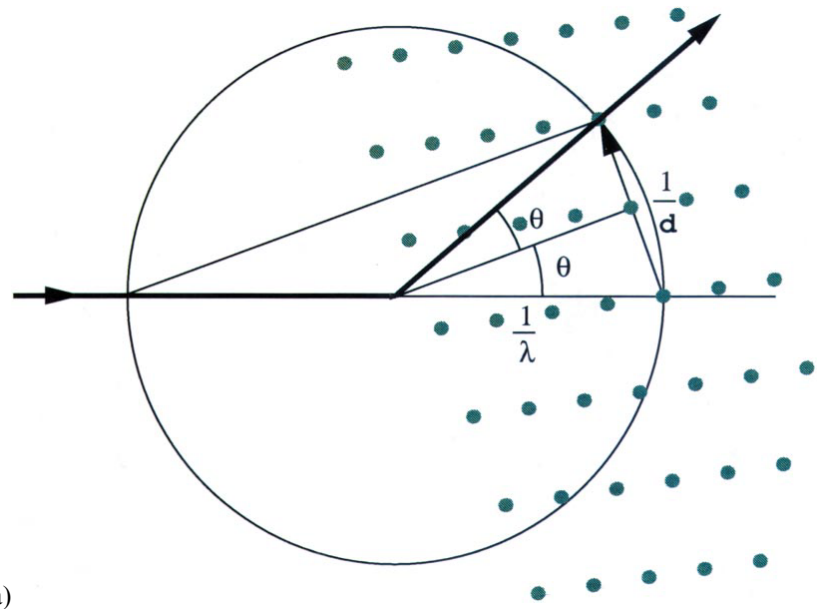
The amplitude of a wave scattered from a particular plane hkl, $|\mathbf{F}_{hkl}|$ is directly quantifiable by the measurement of reflection intensities, I_{hkl}, as $|\mathbf{F}_{hkl}|$ is proportional to the square root of I_{hkl}. However, the phase α_{hkl} can not be measured during the experiment. To achieve the phase information, phasing methods will be explained later in this chapter.

2.3. Data collection

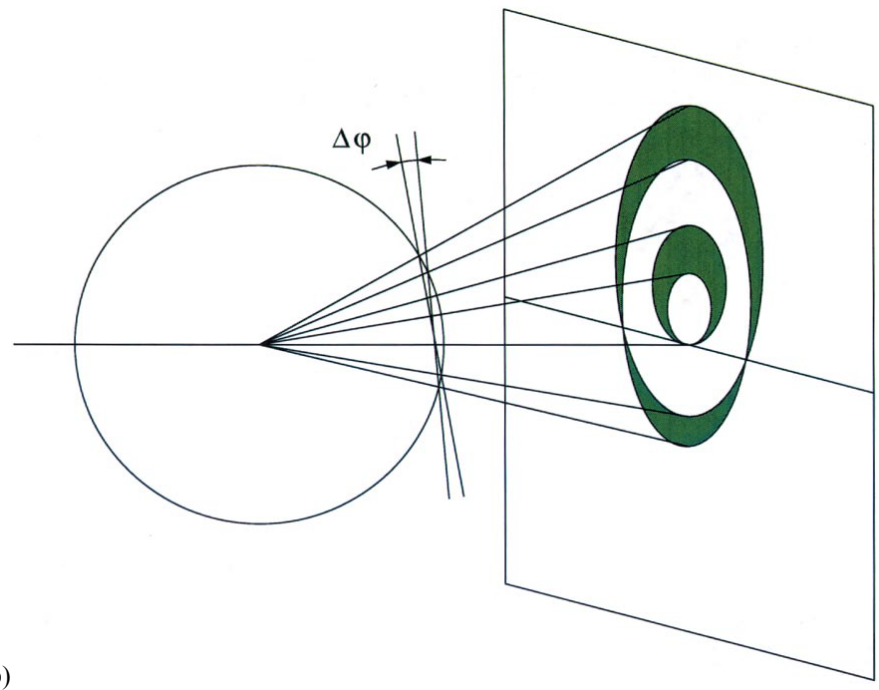
To measure the intensities of all the diffracting rays, a number of images are recorded in which the crystal is rotated through a small angle, perpendicular to the X-ray beam. This method is called the rotation method of data collection (Dauter, 1999) and allows all the data to be measured without individual reflections overlapping with one another.

The radiation of the X-ray that has wavelength λ can be represented by the Ewald sphere with radius $1/\lambda$ centered on the X-ray beam. When reciprocal lattice points with indices (hkl), representing the crystal, lies at the surface of the Ewald sphere, the interference of X-ray scattered for that particular reflection satisfies Bragg's Law in three dimensions. The diffracted beam is directed along the line joining the sphere with the reciprocal lattice point at the surface (Figure 2.5a). To obtain different reflections from other reciprocal lattice points, the reciprocal lattice points need to be moved to the surface of the Ewald sphere. The crystal has therefore to be rotated about a single axis to bring a number of successive reflections into the diffracting position. When the crystal is rotated and the reciprocal plane intersects with the sphere, an ellipse of the diffraction is created by the projection on the detector. The orientation of the plane at the beginning and the ending of the rotation forms two intersecting ellipses with all reflections recorded between them as a lune (Figure 2.5b). There will be a family of concentric lunes displayed on the detector as reflections are arrayed in sets of parallel planes. Diffraction spots are arranged along lines with in each lune, indicating the regularity of the reciprocal lattice. Their pattern is distorted to a

different extent as a result of the mapping of the curved Ewald sphere on the flat detector surface.



a)



b)

Figure 2.5 a) The Ewald construction: $1/d = (1/\lambda)\sin\theta$ is fulfilled when the reciprocal lattice point is on the surface of the Ewald sphere. b) A lune is formed from reflections from the same plane in the reciprocal lattice, corresponding to the start and end positions when the crystal is rotated. This figure was taken from Dauter (1999).

2.3.1. Data collection strategies

In order to carry out an X-ray diffraction experiment, it is necessary to clarify the objectives for each individual crystal. These purposes include sample characterization, structure determination, resolution extension, and ligand identification. Different experimental objectives require different criteria to achieve a good data set. Preliminary X-ray experiments are crucial to analyse the unit cell parameters, diffraction strength and resolution range, as well as to determine space group, however, the complete data collection is not necessarily required at this stage. Different strategies can be used to measure the X-ray diffraction intensities for structure determination, depending on the structure solution method, whether molecular replacement or experimental phasing. Completeness of the data set is prioritised for molecular replacement since the reflection amplitudes at low-resolution ranges are used to calculate Patterson functions. On the other hand, anomalous diffraction phasing requires an accuracy of intensity measurement as well as the data completeness in order to measure the small anomalous signals. A high-resolution data set may be required when atomic structure refinement is performed. In addition, when searching for bound ligands, a modest resolution data set (2.5 Å or better) is sufficient to identify binding without ambiguity.

To obtain the desired data sets, some strategic parameters of data collection are taken into account. A selection of a rotation range and rotation start point is a key to data completeness as the minimum rotation range is required to cover the whole asymmetric unit of reciprocal space which is limited by the resolution sphere and the Laue symmetry group. The orientation of the crystal to the X-ray beam determines where the rotation start point is. Equally, It is important to adjust the rotation per image of the crystal to avoid excessive reflection overlap. $\Delta\varphi$ is generally set to a small increment between 0.1° and 1.5°. The distance between crystal and detector sets the limit of the resolution at the edge of the detector, whilst allowing individual reflections to be separated. Another factor that could affect the data quality is an X-ray exposure time. Long exposure times may lead to overloaded reflections at low resolution as well as radiation damage. For anomalous diffraction experiments, appropriate wavelength(s) used in the

experiment should be selected from the absorption edge of anomalous scatterers, based on the fluorescence spectra.

At present, the strategy generation for data collection is fully automated. However, experimenters need to be able to judge the diffraction quality and decide the practical strategy for meaningful data to be collected.

2.3.2. Radiation damage

During X-ray exposure, more than 90% of the X-ray beam does not interact with the sample at all. A small amount of the X-rays physically interacts with the crystal, defined as diffraction and absorption. Only approximately 8% of these interacting photons are diffracted by the crystal, whereas the other 92% are predominantly absorbed and contribute to the damage of the crystal sample (Garman, 2010). The latter phenomenon involves an energy loss due to either the photoelectric effect or Compton scattering. The photoelectric effect is the total absorption of photons, resulting in an ejection of an inner shell electron of an interacting atom. A primary photoelectron can induce up to 500 ionization incidents which relate to the formation of radical species in the crystal. The excited atom can then release a characteristic X-ray or an Auger electron as it returns to its ground state. Compton scattering represents an inelastic scattering of the photon whose energy is partially transferred to an atomic electron, resulting in lower energy that has higher wavelength. The effect of the ionization of an atom owing to either photoelectric absorption or Compton scattering includes a disruption of crystalline order and specific structural damage. A loss of reflection intensities at high resolution, an expansion of unit cell, an increase in values of the measured internal consistency of data (R_{merge}) and an increase in scaling B-factors values can be observed as a consequence of this global damage. To monitor the global damage during X-ray exposure, the ratio of the summed mean intensity and the mean intensity of the first data set (I_D/I_1) is employed as well as the pairwise R factor between identical and symmetry-related reflections occurring on different images (R_d plot). On the other hand, the specific structural damage is difficult to monitor during the experiment since refined structures are required. Specific structural damage that has been observed in protein structures includes disulfide

bridges broken, the decarboxylation of glutamate and aspartate residues, loss of hydroxyl group of tyrosine residues, C-S bond cleavage in methionine residues plus other covalent bond breakage. Radiation damage often leads to incomplete data collection and failure of structure determination, and must therefore be considered carefully in the design of the experiment.

2.3.3. Tools for data collection

A diffractometer is an instrument designed for collecting X-ray diffraction data. The integral elements of a diffractometer are detailed below:

2.3.3.1. X-ray sources and optics

The two most common ways of producing X-rays for protein crystallography are either by electron bombardment at a target anode or by synchrotron facilities. The generation of X-ray relies on an emission of characteristic radiation which can be selected by a monochromator. The approaching X-ray is diffracted by the crystal monochromator and the desirable wavelength is selected. Focusing optics then allows the X-rays to be focused, into a parallel and intense beam for diffraction experiments.

2.3.3.1.1. Rotating anode

A rotating-anode X-ray generator is commonly set up in many crystallography laboratories. Initially, electrons are accelerated from a heated tungsten filament (cathode) across an electrical potential of several tens of kV towards the target anode. The high energy electron beam strikes the anode and ejects an electron from an inner shell of the target element. An electron from a higher energy shell falls back to fill the empty inner orbital releasing its excess energy as an X-ray photon. The characteristic X-ray radiation from the K shell of a copper anode has an approximate wavelength of 1.54 Å, whereas a shorter wavelength is obtained from molybdenum-based anode (0.71 Å). Due to the large quantity of heat generated by electron attack, the target anode requires sufficient cooling, by rotating the anode. More heat can be dissipated to the water coolant allowing more X-rays to be produced.

2.3.3.1.2. Synchrotron

X-rays can also be generated using a synchrotron and the X-ray beam from synchrotron sources is much more intense and brilliant than the in-house X-ray source. As the energy is tunable, a synchrotron X-ray source has a significant impact on collecting data for some techniques used for determining macromolecular structures. In a synchrotron, electrons are produced in an electron gun, or injector, by thermionic emission and these electrons are then accelerated in a linear accelerator (LINAC) to obtain an extremely high energy. The electron stream then travels to a booster synchrotron in which a radio frequency voltage is used to accelerate the electron energy, whilst being maintained by magnetic field. The final energy obtained is of the order of a few giga electron volts (GeV). The powerful electron beam then travels with a high speed in the storage ring where large electromagnets are installed to bend and keep the electron beam orbit stable in a high vacuum beam tube. Every time the beam is bent, the angular velocity changes and the X-rays are emitted at a tangent to the curved path. The beams thus produced are called bending magnet beams. Alternatively, insertion devices such as wigglers or undulators can be placed in the straight sections of the storage ring. These insertion devices are a series of powerful magnets which oscillate the beam, rapidly produce even more brilliant X-rays. The X-ray beams pass through a front end, to a beam-line where an experimental hutch is established for carrying out the X-ray diffraction experiment.

2.3.3.2. Detectors

Nowadays, the use of conventional X-ray film for collecting X-ray diffraction images has been replaced by imaging plate and charged coupled device (CCD) detectors greatly increasing the speed of data collection (Smyth & Martin, 2000).

Imaging plate detectors consist of a layer of phosphor BaFBr doped with Eu^{2+} based material. The incident X-ray stimulates the Europium ions to their excited state (Eu^{3+}). After scanning by a He-Ne laser, phosphor-luminescence is induced and photons are released. The emitted light intensities are recorded by a photo multiplier tube and thus an image of the diffraction pattern can be reconstructed. The image plate is erased after each exposure by visible light exposure after readout.

Unlike imaging detectors, CCD detectors are technologically advanced and more favourable in terms of fast readout time, therefore they are predominately established in many synchrotron macromolecular crystallography beamlines. CCD detectors are made of CCD semiconductor array chips. X-ray photons are absorbed in a thin fluorescent screen coated with phosphor Gd_2O_2S doped with Tb which converts X-ray to visible light photons and generates scintillation. The screen is attached to fiber optic taper next to a CCD semiconductor in which electrons are accumulated proportional to the number of photons before being relocated into the readout row and producing electronic diffraction pattern images.

Recently, solid state detectors, such as the hybrid pixel detector PILATUS 1M have been introduced. These detectors record the X-rays directly onto each pixel, allowing continuous read out and very fast data collection times.

2.3.3.3. Goniometer

To be able to rotate the crystal, the apparatus called goniometer or goniostat is employed. A loop with a cryoprotected crystal is mounted on a goniometer head that allows the crystal movement and automatic adjustment. A rotation of the goniometer, φ-angle, is assigned the crystal orientation with reference to the X-ray beam and the detector while maintaining the crystal centering. In addition, heating the goniometer head is required in order to prevent ice formation from the crystal cooling system.

As the majority of the incident X-rays are not scattered by the electrons in the crystal, a beam stop, made from a tiny lead piece is placed between the mounted crystal and the detector. This object prevents an overexposure on the detector from the intense direct beam which may otherwise damage it.

2.3.3.4. Cryocooler

The cryocooling system provides a cryogenic nitrogen stream over the crystal during an X-ray exposure to reduce radiation damage. In the cryocooler, liquid nitrogen from a dewar is first evaporated and the subsequent cooling of the gas to 100K is undertaken by heat exchange with the liquid nitrogen. The cold stream of

cryogenic nitrogen gas is gently blown out over the crystal via the nozzle and is surrounded by a dry air stream which counteracts the formation of frost on the crystal.

2.4. Data processing

All reflections produced from a diffracting single crystal need to be processed in order to obtain a useful dataset for subsequent phasing and electron density reconstruction. This stage involves the determination of unit cell dimensions, crystal system and space group, data integration, intensity measurement, scaling and averaging intensities and calculating structure factor amplitudes. The data processing is done automatically by software packages and program suites.

2.4.1. Autoindexing

Using a couple of diffraction images, usually 90° apart from each other, the program Mosflm (Leslie & Powell, 2007) will find diffraction spots that have higher intensity pixels than background. The identification of these reflections enables the program to index the crystal and obtain unit cell dimensions by measuring the reflection spacings which are proportional to reciprocal lattice spacings and geometrically related to real lattice spacings. To determine the unit cell dimensions, the crystal-to-detector distance must be known as well as the direct beam position and the incident wavelength. Therefore, the unit cell dimensions, crystal system and lattice type can be computed and a list of possible solutions with their degrees of penalty is produced.

2.4.2. Integration

In order to obtain a single data set, all images collected from the X-ray diffraction experiment are integrated. This process involves predicting the position of the reflections and estimating the reflection intensities with associated standard deviations. The intensities calculated by a summation integration and a subsequent profile fitting. The pixel values of all intensity pixels are summed and the sum of background values are subtracted. The standard deviation is based on Poisson statistics. During the integration of the complete dataset, the unit cell parameters are typically fixed but the detector parameters, crystal orientation and

mosaic spread are refined to establish the best prediction of spot positions. The data integration results are written in a single MTZ file including the indices of all reflections (h,k,l) and estimated intensities (I_{hkl}) and error σ(I).

2.4.3. Data reduction

To reduce lists of reflections to a unique list, the integrated raw data are scaled and the multiple measurements of identical reflections are merged. Scaling is required to account for variations in incident beam intensity, crystal shape and absorption. The average signal to noise ratio over all N reflections in a resolution shell can be expressed as:

$$\frac{I}{\sigma_{(I)}} = \frac{1}{N}\sum_{h}\frac{|I_{(h)}|}{\sigma_{(I_{(h)})}}$$

Symmetry-related reflections are merged into a unique data set representing the asymmetric unit of the reciprocal space. Merging statistics are computed to indicate the data quality in terms of the reliability factor. The linear merging R-value is commonly used to describe the merging of N redundant observations of reflections h within the desired resolution range.

$$R_{merge} = \frac{\sum_h \sum_{i=1}^{N} |I_{(h)i} - \bar{I}_{(h)}|}{\sum_h \sum_{i=1}^{N} I_{(h)i}}$$

where $\bar{I}_{(h)}$ is the average intensity of each reflection

As the number of equivalent measurements increases, so does R-merge and so a redundancy-independent merging R-value can also be calculated.

$$R_{rim} \text{ or } R_{meas} = \frac{\sum_h (\frac{N}{N-1})^{\frac{1}{2}} \sum_{i=1}^{N} |I_{(h)i} - \bar{I}_{(h)}|}{\sum_h \sum_{i=1}^{N} I_{(h)i}}$$

Introducing a term of $1/(N-1)^{1/2}$ to the linear merging R-value can improve the precision of merging intensities, yielding the precision-indicating merging R-value.

$$R_{pim} = \frac{\sum_h (\frac{1}{N-1})^{\frac{1}{2}} \sum_{i=1}^{N} |I_{(h)i} - \bar{I}_{(h)}|}{\sum_h \sum_{i=1}^{N} I_{(h)i}}$$

2.4.4. Determining data quality

At this stage, a unique dataset has been obtained and an initial analysis of the quality of diffraction data is carried out. There are several criteria to assess the data quality. Signal-to-noise ratio (SNR) is one of data quality indicators, either the overall SNR of the data or the SNR in the highest resolution shell. $1 < I/\sigma(I) < 2$ is typically defined as a SNR threshold for the resolution cut-off. The completeness of good data is expected to be close to 100% across the entire resolution range. Merging statistics are other important indicators to describe the internal consistency of the data as well as indicate the level of random error and residual systematic error in the data (Evans *et al*, 2011).

2.5. Structure determination

At this stage, the intensity of every reflection has been measured (I_{hkl}), which is proportional to $|\mathbf{F}_{hkl}|^2$, the remaining unknown is the phase α_{hkl}. As this phase information has been lost during the X-ray diffraction experiment, solving the phase problem is required in order to fulfill the electron density reconstruction according to the equation:

$$\rho(x,y,z) = \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}| \; e^{i\alpha} e^{-2\pi i(hx+ky+lz)}$$

There are two possible ways for phasing that are molecular replacement and experimental phasing. Providing that there are closely related protein structures available in the databases, molecular replacement can be simply employed. If this is not the case, then experimental phases must be obtained. This method involves

finding the substructure of either heavy atoms or anomalous scattering atoms existing in the protein molecules with different techniques subsequently used to derive the phase angle. The common methods used to determine the phase angle are described below, and both rely on the Patterson function to determine the heavy atom substructure.

2.5.1. Patterson function

As the phase of an atomic structure factor depends on the position of the atom in the unit cell, to find the coordinates of the heavy atom in the unit cell, the Patterson method is employed. The Patterson function P (u,v,w) is a Fourier sum of the square of the structure factor amplitude with its frequencies h in the u-direction, k in the v-direction and l in the w-direction, in the absence of phases.

$$\mathrm{P(u,v,w)} = \frac{1}{\mathrm{V}} \sum_{\mathrm{h}} \sum_{\mathrm{k}} \sum_{\mathrm{l}} |\mathbf{F}_{\mathrm{hkl}}|^2 \; \mathrm{e}^{-2\pi \mathrm{i}(\mathrm{hu+kv+lw})}$$

The amplitude contribution of the heavy atom in the derivative crystal can be obtained from the difference between the structure-factor amplitudes of the native and the derivative.

$$(\Delta\mathbf{F})^2 = (|\mathbf{F}_{\mathrm{PH}}| - |\mathbf{F}_{\mathrm{P}}|)^2$$

A difference Patterson function can be derived as the equation below

$$\Delta\mathrm{P(u,v,w)} = \frac{1}{\mathrm{V}} \sum_{\mathrm{h}} \sum_{\mathrm{k}} \sum_{\mathrm{l}} (\Delta\mathbf{F}_{\mathrm{hkl}})^2 \; \mathrm{e}^{-2\pi \mathrm{i}(\mathrm{hu+kv+lw})}$$

A Patterson map shows peaks at the locations corresponding to vectors between atoms and allows location of the atom(s) in the unit cell.

2.5.2. Isomorphous replacement

Due to the isomorphism of native and heavy-atom derivative crystals, structure factors of a single reflection from native and the corresponding derivative reflection can be used to find out the heavy atom contribution. In vector terms, the structure factor for the heavy atom derivative ($\mathbf{F}_{PH}$) is the sum of those for the native protein ($\mathbf{F}_P$) and the heavy atom alone ($\mathbf{F}_H$) as the following:

$$\mathbf{F}_{PH} = \mathbf{F}_P + \mathbf{F}_H$$

or

$$\mathbf{F}_P = \mathbf{F}_{PH} - \mathbf{F}_H$$

The structure factor amplitude of the heavy atoms can be estimated from the isomorphous difference.

$$|\mathbf{F}_H| \simeq |\mathbf{F}_{PH}| - |\mathbf{F}_P|$$

If the substructure of the heavy atoms can be determined using the Patterson function, then $\mathbf{F}_H$ can be calculated and $|\mathbf{F}_H|$ and α_H are thus known.

This vector equation can be solved for $\mathbf{F}_P$ using a Harker diagram which represents the relative vectors in the complex plane (Figure 2.6a). From the measured reflection intensities of the protein and heavy atom derivative crystals, their structure factor amplitudes $|\mathbf{F}_P|$ and $|\mathbf{F}_{PH}|$ can be calculated as the length of the vectors $\mathbf{F}_P$ and $\mathbf{F}_{PH}$ without their direction or phase angles. In the Harker diagram, a circle of radius $|\mathbf{F}_{PH}|$ centered on the head of the vector $-\mathbf{F}_H$ is drawn, representing the vector sum $|\mathbf{F}_{PH}| - \mathbf{F}_H$ and together with a circle of radius $|\mathbf{F}_P|$ centered at the origin. Therefore, the possible vector solutions can be drawn at the intersection points of these two circles (Figure 2.6b). To break the phase ambiguity, the second heavy atom derivative is used to obtaining the more likely phase angle from the overlapping of three circles (Figure2.6c). The phase of the native protein α_P can be calculated from the following equation.

$$\alpha_P = \alpha_H \pm \cos^{-1}\left[\frac{(\mathbf{F}_{PH}^2 - \mathbf{F}_P^2 - \mathbf{F}_H^2)}{2\mathbf{F}_P\mathbf{F}_H}\right]$$

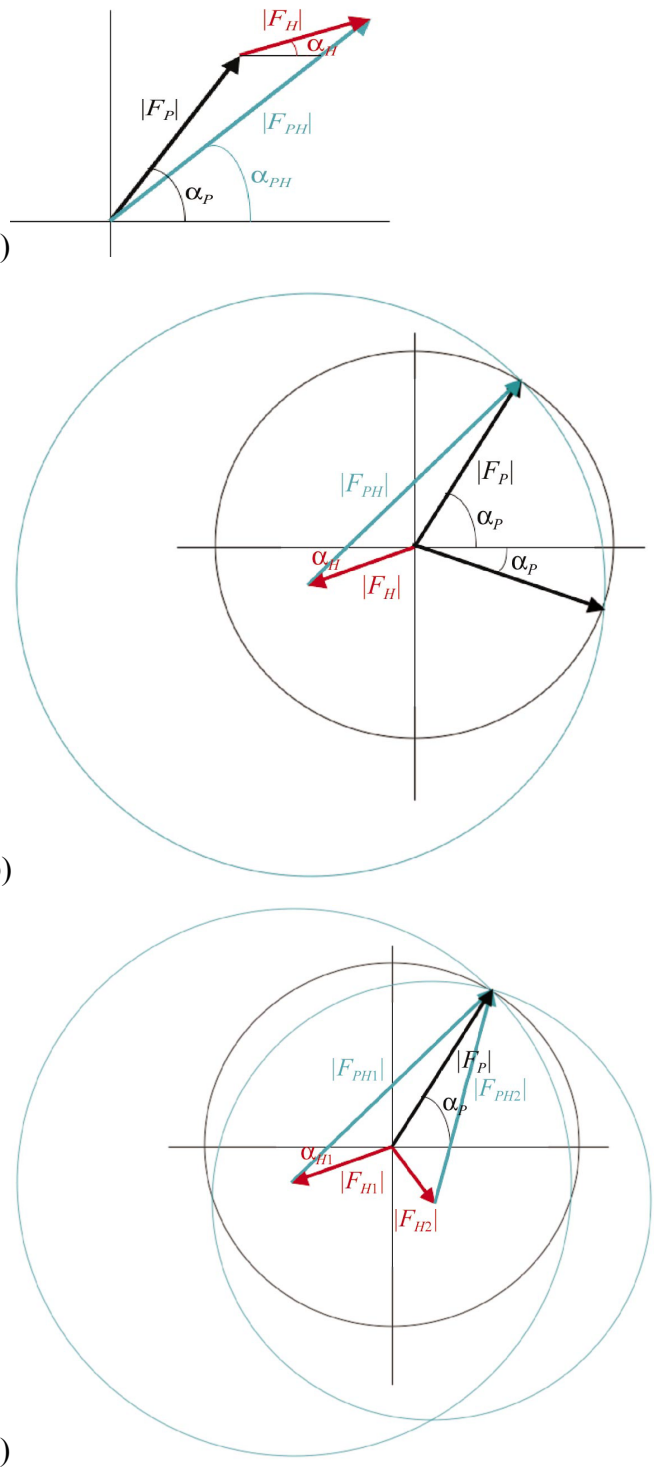


Figure 2.6 Isomorphous replacement a) A representation of structure factor vectors of a native protein $\mathbf{F}_P$ with phase α_P (black) and its heavy atom derivative $\mathbf{F}_{PH}$ with phase α_{PH} (blue) is shown in an Argand diagram. b) Harker construction of single isomorphous replacement with phase ambiguity: two possible phases (α_P) can be calculated. c) Harker construction of multiple isomorphous replacement: two or more heavy atom derivatives are used to resolve the phase ambiguity. The illustrations are taken from Taylor (2010).

2.5.3. Multi-wavelength anomalous dispersion (MAD)

A different way to obtain phases is to use the anomalous scattering signal from certain atoms in the structure. In terms of X-ray absorption, a group of particular atoms can exhibit anomalous scattering when the incident wavelength is near to their absorption edges in which the absorption dramatically changes. The anomalous differences can be employed to locate the anomalous scatterers.

Selenium is the most widely used as an anomalous scattering atom for MAD phasing approach. A fluorescence scan close to the absorption of the Se K-edge (0.9795 Å) where the f' and f'' are greatest allows the identification of energies to be used in MAD data collection (Figure 2.7). Three wavelengths are selected based on the absorption curve. λ_1 is at the peak of the anomalous signal f''. λ_2 is at the point of inflection where the dispersive signal f' reaches its minimum. λ_3 is a remote wavelength that maximises the dispersive difference to λ_2.

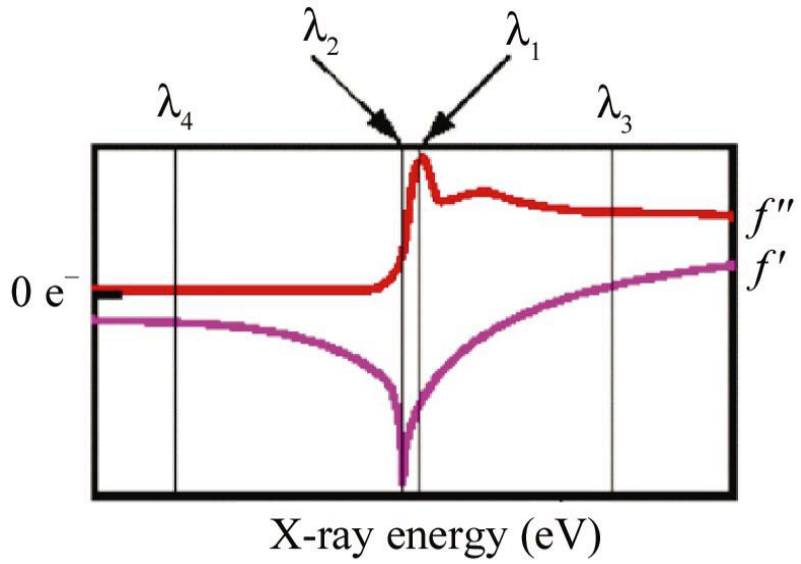


Figure 2.7 Typical absorption curve showing anomalous scattering signal in electrons for MAD phasing: The selection of wavelengths used in MAD is based on the real and imaginary anomalous scattering contributions, f' (purple) and f'' (red). This illustration is taken from Taylor (2010).

An atomic scattering factor relating to the anomalous contributions can be described as:

$$f_{(\theta,\lambda)} = f^{0}_{(\theta)} + f'_{(\lambda)} + i \cdot f''_{(\lambda)}$$

The atomic scattering factor involves three different terms in which f^0 represents a normal scattering depending on the Bragg angle whereas the dispersive component f' and the absorption component f'' represent anomalous scattering occurring at the absorption edge depending on the wavelength. The normal scattering factor is modified by f' while f'' is 90° advanced in phase. This anomalous effect leads to the breakdown of the Friedel's law (Figure 2.8). As a result, the structure factor amplitude and intensity of reflection hkl is not equal to -h-k-l. The Bijvoet pair $\mathbf{F}^+$ and $\mathbf{F}^-$ can be expressed in the following equations.

$$|\mathbf{F}^+|^2 = |\mathbf{F}_\mathrm{T}|^2 + \mathrm{a}|\mathbf{F}_\mathrm{A}|^2 + \mathrm{b}|\mathbf{F}_\mathrm{T}||\mathbf{F}_\mathrm{A}| \cos\alpha + \mathrm{b}|\mathbf{F}_\mathrm{T}||\mathbf{F}_\mathrm{A}| \sin\alpha$$

and

$$|\mathbf{F}^-|^2 = |\mathbf{F}_\mathrm{T}|^2 + \mathrm{a}|\mathbf{F}_\mathrm{A}|^2 + \mathrm{b}|\mathbf{F}_\mathrm{T}||\mathbf{F}_\mathrm{A}| \cos\alpha - \mathrm{b}|\mathbf{F}_\mathrm{T}||\mathbf{F}_\mathrm{A}| \sin\alpha$$

where $|\mathbf{F}_\mathrm{T}|$ is the total structure factor amplitude of the protein without anomalous contributions and $|\mathbf{F}_\mathrm{A}|$ is the structure factor amplitudes of the anomalous marker. The constants a, b and c are dependent on the anomalous scattering contributions as described below.

$$a = \frac{f'^2 + f''^2}{f^{0\,2}} \qquad b = \frac{2f'}{f^0} \qquad c = \frac{2f''}{f^0}$$

The phase angle α is the phase difference between the true protein phase φ_T and the phase of the anomalous scattering contributions φ_A (Figure 2.9).

$$\alpha = \varphi_\mathrm{T} - \varphi_\mathrm{A}$$

When data collected at more than one wavelength are available, these unknown variables can be determined and $\mathbf{F}_\mathrm{A}$, $\mathbf{F}_\mathrm{T}$ and α can be computed.

From the broken Friedel's law, the Bijvoet difference can be calculated as:

$$\Delta \mathbf{F}^{\pm} = |\mathbf{F}_{\mathrm{PA}}^{+}| - |\mathbf{F}_{\mathrm{PA}}^{-}|$$

To locate anomalous scattering atoms, the Bijvoet difference coefficients $\Delta \mathbf{F}^2$ are used to calculate a Patterson map that should contain only peaks corresponding to the interatomic vectors between pairs of anomalous scattering atoms.

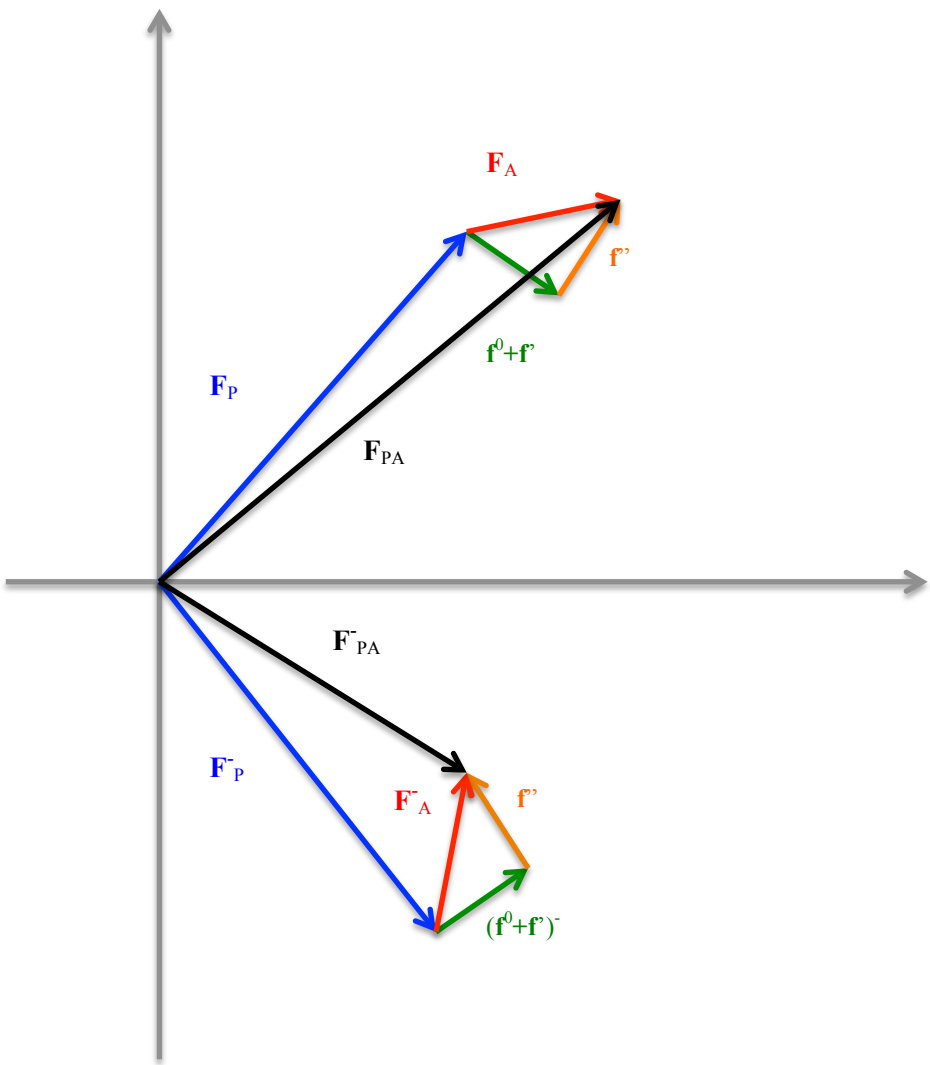
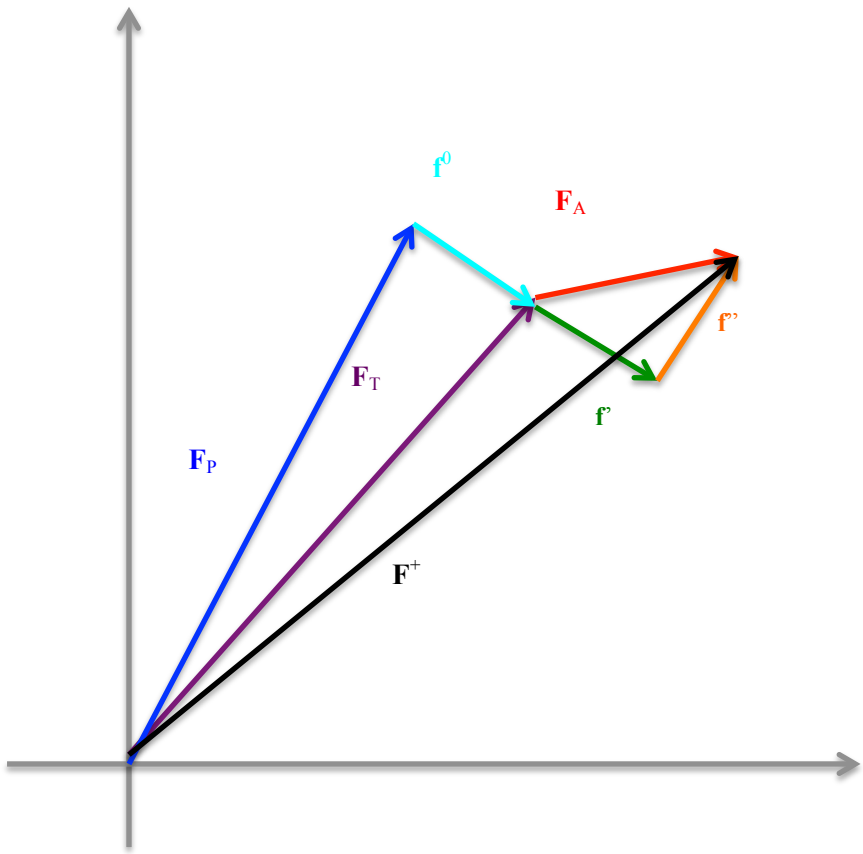
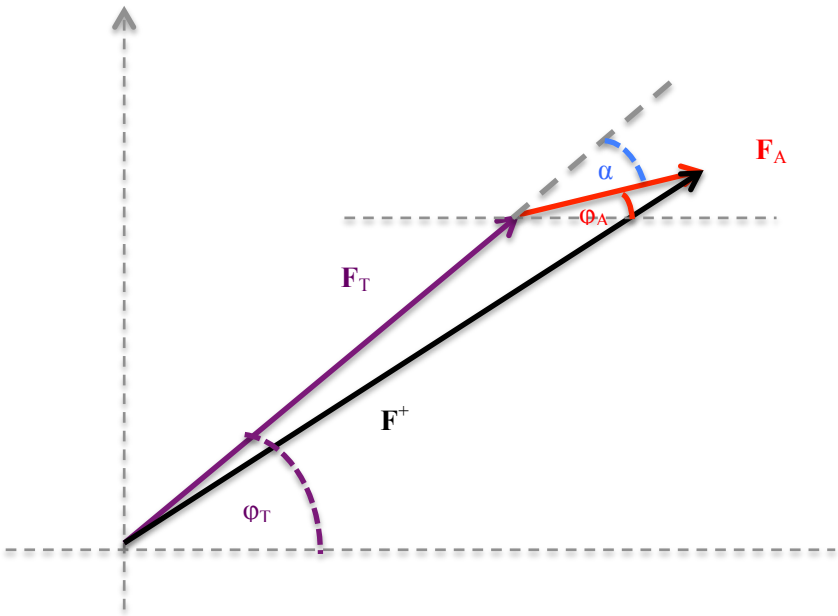


Figure 2.8 Breakdown of Friedel's law: When an anomalous scatterer is present, the structure factor amplitudes of Bijvoet mates are different, $|\mathbf{F}_{\mathrm{PA}}| \neq |\mathbf{F}^{-}_{\mathrm{PA}}|$. The vector $\mathbf{F}_{\mathrm{P}}$ (blue) represents the contributions from all of non-anomalous atoms and the vector $\mathbf{F}_{\mathrm{A}}$ (red) is the contributions from the anomalous scattering atom, consisting f^0+f' (green) and f'' (orange) components.



a)



b)

Figure 2.9 Argand diagrams involved in MAD phasing: a) The structure factor of the protein with the anomalous scattering atom $\mathbf{F}^+$ (black) is made of the contribution of the protein $\mathbf{F}_T$ (purple) and the anomalous scattering contributions $\mathbf{F}_A$ (red). b) The phase shift (α) is described as the difference between the protein phase without anomalous contributions (φ_T) and the phase of the anomalous scattering contributions (φ_A).

2.5.4. Molecular replacement

If the protein of interest has a structurally similar model available, phasing can be achieved by using the phase from its structure homologue. Typically, more than 25% sequence identity and less than 2.0 Å rmsd between Cα atoms of the known models and unknown structured protein are considered to be required for this technique to work. The principle of molecular replacement based structure solution relies on the Patterson function. A Patterson map is calculated using a Fourier transform of the reflection intensities, resulting in an interatomic vector map. A Patterson map is also generated using calculated structure factor amplitudes from the atomic coordinates of the known structure model. In order to solve the unknown structure, the Patterson functions of the model and the unknown protein structures are aligned. The Patterson map of the known structure model is rotated over the other to obtain the relative orientation of the search model in the new unit cell. A new Patterson is then calculated on the related search model and the translation of the correctly oriented model is obtained by searching similar Patterson vectors between symmetry-related molecules in the unit cell. Therefore, an initial electron density map of the target protein structure can be reconstructed using its own measured amplitudes and phases taken from the correctly oriented search model.

2.6. Structure refinement

After an initial model of the protein has been built into the electron density map, model refinement is carried out to make the best agreement between the model and the X-ray data. A local real space refinement is approached by fitting individual residues into the electron density. To get more favorable geometry, the atomic positions are adjusted using real space regularization tools. After the local real space refinement, stereochemical and conformational errors remain. The model is therefore subjected to global reciprocal space restrained refinement. Stereochemical restraints including bond length, bond angle, torsion angle, chirality and planarity are taken into account during refinement, which has the effect of improving the observation-to-parameter ratio of the refinement. A number of repeating cycles of refinement are performed to adjust the atomic parameters of the model and optimise the fit between observed and calculated

structure factor amplitudes until the shifts reach convergence. An overall fitness between diffraction data and the model structure is described as a global linear residual, R-value.

$$\mathrm{R} = \frac{\sum_h |\mathbf{F}_{\mathrm{obs}} - \mathbf{F}_{\mathrm{calc}}|}{\sum_h \mathbf{F}_{\mathrm{obs}}}$$

The agreement between the experimental data and fitted model can be computed separately for the working data set and the excluded cross validation data set, designated as R_{work} and R_{free}, respectively. A small subset of randomly selected reflections set aside and not involved in the refinement is used for cross-validation technique to represent how well the model predicts experimental observations that are not themselves used to fit the model. Thus, R_{free} is used to evaluate the model quality as well as the phase accuracy (Brunger, 1992).

2.6.1. Maximum likelihood

Maximum likelihood is statistical inference technique that declares that the best model on the evidence of the data is the one that explain what has actually been observed with the highest probability (McCoy, 2004). The maximum likelihood methods play an important role in crystallographic programme algorithms involving in data processing, electron density reconstruction and structure refinement. In protein crystallography, an experimental data refers to the observed structure factor amplitude for each reflection in reciprocal space ($|\mathbf{F}_{obs}|$) and a model means the calculated structure factor of the structure in real space ($\mathbf{F}_{cal}$). A solution is the most likely model with the $\mathbf{F}_{cal}$ that has the highest probability of generating the $\mathbf{F}_{obs}$. A reflection likelihood function can be written as:

$$\mathrm{P} - Xray = \mathrm{P}(data; model) = \mathrm{P}(|\mathbf{F}_{\mathrm{obs}}|; |\mathbf{F}_{\mathrm{cal}}|)$$

As the assumption of an independence of all reflections, an approximation of total likelihood is the product of the reflection likelihoods.

$$P - Xray_{total} = \prod_{h} \mathrm{P}(|\mathbf{F}_{\mathrm{obs}}|; |\mathbf{F}_{\mathrm{cal}}|) = \sum_{h} \mathrm{P}(|\mathbf{F}_{\mathrm{obs}}|; |\mathbf{F}_{\mathrm{cal}}|)$$

Since the probability distributions are continuous, the sum in this equation is converted into an integral with all angles.

$$P - Xray_{total} = \int_{0}^{2\pi} \mathrm{P}(|\mathbf{F}_{\mathrm{obs}}|; |\mathbf{F}_{\mathrm{cal}}|)\, dx$$

The likelihood function is given by integrating out the nuisance phase between $\mathbf{F}_{\mathrm{obs}}$ and $\mathbf{F}_{\mathrm{cal}}$, α.

$$\mathrm{P}(|\mathbf{F}_{\mathrm{obs}}|; |\mathbf{F}_{\mathrm{cal}}|) = \int_{0}^{2\pi} \mathrm{P}(|\mathbf{F}_{\mathrm{obs}}|, \alpha; |\mathbf{F}_{\mathrm{cal}}|)\, dx$$

The errors in the phased structure factors in reciprocal space that arises from errors in the atomic model in real space can be described by the central limit theorem. The probability distribution for $\mathbf{F}_{\mathrm{obs}}$ given $\mathbf{F}_{\mathrm{cal}}$ is a two dimensional Guassian in reciprocal space with variance σ^2_{Δ} centered on $\mathbf{F}_{\mathrm{cal}}$.

$$\mathrm{P}(\mathbf{F}_{\mathrm{obs}}; \mathbf{F}_{\mathrm{cal}}) = \frac{1}{\pi\sigma_{\Delta}^2}\, e^{-\frac{|\mathbf{F}_{\mathrm{obs}} - \mathbf{F}_{\mathrm{cal}}|^2}{\sigma_{\Delta}^2}}$$

From the relationship, $\mathrm{P}(|\mathbf{F}_{\mathrm{obs}}|,\alpha;|\mathbf{F}_{\mathrm{cal}}|) = \mathbf{F}_{\mathrm{obs}} \times \mathrm{P}(\mathbf{F}_{\mathrm{obs}};\mathbf{F}_{\mathrm{cal}})$, it is therefore,

$$\mathrm{P}(|\mathbf{F}_{\mathrm{obs}}|; |\mathbf{F}_{\mathrm{cal}}|) = \frac{2|\mathbf{F}_{\mathrm{obs}}|}{\pi\sigma_{\Delta}^2} \exp\left(-\frac{|\mathbf{F}_{\mathrm{obs}}|^2 + |\mathbf{F}_{\mathrm{cal}}|^2}{\sigma_{\Delta}^2}\right) \int_{0}^{2\pi} \exp\left(\frac{2|\mathbf{F}_{\mathrm{obs}}||\mathbf{F}_{\mathrm{cal}}|}{\sigma_{\Delta}^2}\, cos\alpha\right) d\alpha$$

The equation is thus simplified to the Rice distribution equation,

$$\mathrm{P}(|\mathbf{F}_{\mathrm{obs}}|; |\mathbf{F}_{\mathrm{cal}}|) = \frac{2|\mathbf{F}_{\mathrm{obs}}|}{\sigma_{\Delta}^2} \exp\left(-\frac{|\mathbf{F}_{\mathrm{obs}}|^2 + |\mathbf{F}_{\mathrm{cal}}|^2}{\sigma_{\Delta}^2}\right) I_0\left(\frac{2|\mathbf{F}_{\mathrm{obs}}||\mathbf{F}_{\mathrm{cal}}|}{\sigma_{\Delta}^2}\right)$$

When $|\mathbf{F}_{cal}|$ is equal to zero, the Wilson distribution equation is therefore introduced,

$$\mathrm{P}(|\mathbf{F}_{\mathrm{obs}}|; |\mathbf{F}_{\mathrm{cal}}| = 0) = \frac{2|\mathbf{F}_{\mathrm{obs}}|}{\sigma_{\Delta}^{2}} \exp\left(\frac{|\mathbf{F}_{\mathrm{obs}}|^{2}}{\sigma_{\Delta}^{2}}\right)$$

Usually, the log-likelihood (LL) is calculated instead of the likelihood. The LL gain value, which is the difference between LLs, can be used to determine the success of a procedure in maximum likelihood-based programmes.

2.7. Structure validation

Structure validation is an assessment of geometry and electron density to ensure the reliability of the structure model. The model can be checked automatically using structure validation tools such as PROCHECK (Laskowski *et al*, 1993) and MolProbity (Chen *et al*, 2010). The stereochemical parameters of the protein structure are verified in terms of an overall goodness of stereochemistry G-factor. G-factors should be scored greater than -0.5. The Ramachandran plot is used to evaluate the backbone conformation of the protein (Ramachandran *et al*, 1963). All residues should be in the allowed regions of the Ramachandran plot. If there is wrong conformation within the structure, it will have Ramachandran outiers. The distribution plot of B-factors can reveal an incorrection of model building either main chains or side chains with an observation of high B-factor values.

As refinement proceeds, the model is validated using these tools, and corrections are made to areas with bad geometry, high B-factors, incorrect torsion angles and Ramachandran outliers. These tools are also useful to check that the final model contains no errors and is of a sufficient quality for deposition in the protein data bank.

Chapter 3: Materials and Methods

This chapter describes the materials, methods and strategies used throughout the research project. Target selection, the amplification of the protein encoded genes, cloning of these genes, protein expression and purification, protein crystallisation, X-ray data collection, data processing, structure determination and ligand observation by nuclear magnetic resonance (NMR) are included.

3.1. Target selection

All target genes in this project were selected from lists of identified potential targets based on literature reviews (see section 1.8). As there is increasing attention on the development of rational drug design and biomarkers, putative essential genes and also immunogenic proteins of *B. pseudomallei* were taken into account in preparing the target list. Seven *B. pseudomallei* genes of interest are described in section 4.1.

3.2. DNA manipulation

3.2.1. Gene amplification

The genes encoding target proteins were amplified from the genomic DNA of *B. pseudomallei* strain D286 isolated from a melioidosis-suffering patient in Malaysia. Oligonucleotide primers were designed using nucleotide sequences of the completely sequenced genome of *B. pseudomallei* strain K96243 as a reference. They were chemically synthesised by commercial company services (Eurofins MWG Operon and Sigma-Aldrich) for subsequent polymerase chain reaction (PCR) according to published guidelines (Dieffenbach *et al*, 1993). The standard 3-step PCR was carried out using a thermal cycler TC-312 (Techne). The reaction was made up of *B. pseudomallei* genomic DNA template, primers, polymerase enzyme, buffer and deoxyribonucleotides (dNTPs). Two different protocols were employed depending on the cloning strategies (see Table 3.1). Protocol A was used with Biomix™ Red (Bioline) for TA cloning whereas DyNAzyme™ EXT DNA polymerase (Finnzymes) was used together with its

optimised buffer and dNTPs for restriction enzyme cloning (protocol B). To make the DNA template accessible, the initial denaturation at 94˚C was performed for 5 minutes. In each cycle, denaturation at 94˚C for 1 minute was carried out and the reaction was then cooled down to the annealing temperature calculated on the basis of the melting temperatures of the oligonucleotide primers (Tm) for 1 minute. Subsequently, the extension of DNA fragments was followed at 72˚C for 3 minutes. 25 cycles were repeated and the final extension occurred at 72 ˚C for 5 minutes as the final step.

Table 3.1 Recipe for PCR

	Protocol A (µl)	Protocol B (µl)
Genomic DNA of *B. pseudomallei*	1	1
100 µM forward primer	0.5	0.5
100 µM reverse primer	0.5	0.5
2x Biomix Red	25	-
DyNAzyme™ EXT DNA Polymerase	-	1
10x DyNAzyme™ EXT Buffer	-	5
10 mM dNTP mix	-	1
Optional DMSO	ranging from 0-1 µl (up to 10% v/v)	
Sterile deionised water	to give final volume of 50 µl	

3.2.2. Agarose gel electrophoresis

To analyse DNA fragments including PCR products and digestion products, 1% (W/V) agarose gels were prepared in 50 ml 1x TAE buffer and 4 µl GelRed™ (Biotium). 5-10 µl DNA samples with an addition of 6x gel loading dye (NEB) were run on the gel in TAE buffer at 100 Volts for 40 minutes. To estimate the size of the DNA fragments, an appropriate molecular weight DNA marker (HyperLadder™ I, Bioline) was used on the same gel. The gel containing DNA samples was visualised under ultra-violet light.

3.2.3. PCR purification and gel extraction

After PCR, the PCR products were monitored on a 1% agarose gel and if the specific PCR product had been obtained, then it was separated from the rest of the reaction mixture using the QIAquick PCR Purification Kit (QIAGEN) according to the manufacturer's instructions. When non-specific PCR products had been seen on the agarose gel, the rest of the PCR mixture was then loaded and run on another agarose gel electrophoresis. The specific DNA band with an expected size of the gene was cut off and recovered from the gel using a QIAquick Gel Extraction Kit (QIAGEN) following the manufacturer's instructions.

3.2.4. Gene cloning

After PCR purification or gel extraction, PCR products were cloned into pET vectors following the manufacturers' protocols.

3.2.4.1. TA cloning

Most of target genes (including BPSS0945, BPSS0238, BPSS1416, BPSS0603 and BPSS0683) in this project were engineered into the pETBlue-1™ AccepTor vector (Novagen) which provides a high copy number of the screening vector for cloning and also facilitates protein expression (Figure 3.1 in Appendix). As a result of PCR, the target genes were synthesised with 3'-deoxyadenosine triphosphate overhangs, which were ligated to 3'-deoxyuridine triphosphate overhangs present in the linearised vector. With the AccepTor vector kit (Novagen), the PCR product was gently mixed with the commercially linearised vector, which was then ligated by the ligase provided in the Clonables™ ligation premix. The recipe for ligation is shown in Table 3.2. The ligation reaction was performed at 16°C for 2 hours.

Table 3.2 A ligation protocol for the pETBlue-1™ AccepTor vector (Novagen)

	µl
PCR product	4
Clonables™ 2x Ligation Premix	5
50 ng/µl AccepTor™ Vector	1
Total reaction volume	10

3.2.4.2. Restriction enzyme cloning

Some of target genes in this work (including truncated BPSL3022 and BPSL0606) were cloned into the pET24-a vector (Figure 3.2 in Appendix) using restriction enzymes. Firstly, the gene sequences were analysed for potential cleavage sites using NEBcutter (Vincze *et al*, 2003), according to the recognition sequences of the commercially available restriction enzymes. Restriction enzymes that would not digest the gene targets were considered to be used for cloning, corresponding to the restriction map of the vector.

The vector was digested with two different restriction enzymes for example *Nde*I and *Bam*HI (*Eco*RI or *Xho*I are used instead of *Bam*HI for cloning BPSL3022). To prepare the linearised vector ready to be used for cloning, the digestion mixture is shown below (Table 3.3).

Table 3.3 Restriction enzyme digestion

	µl
pET24-a vector	10
*Nde*I	1
*Bam*HI	1
10x buffer*	2
Sterile deionised water	6
Total reaction volume	20

*The selection of buffer depends on individual enzyme activity efficiency. The compatible buffer for both enzymes is recommended by the manufacturer (NEB).

The digestion was carried out at 37 °C, which is an optimal temperature for enzyme activity, for 1 hour. Then, the linearised vector was recovered by the PCR purification kit (QIAGEN). The gene fragment from the PCR reaction was prepared to be an overhang-provided insert. The digestion was performed in the same way as the vector preparation.

Digested vector and insert were mixed together with T4 ligase to join each other with compatible sticky ends as the recipe in Table 3.4. The ligation reaction was incubated at 16 °C for 3 hours or overnight for enhancing recombinant clones.

Table 3.4 Ligation for recombinant pET24-a

	µl
Digested pET24-a	5
Digested gene insert	12
T4 Ligase	1
10x T4 ligase buffer	2
Total reaction volume	20

3.2.5. Transformation of *E. coli* competent cells

After cloning, the engineered pETBlue-1 vectors were transformed into NovaBlue (Novagen), a derived K12 strain of *E. coli* recommended for the pETBlue-1 system.

1 µl of ligation mixture was added to 50 µl of competent cells and placed on ice for 30 minutes. Using heat shock to allow physical changes in the bacterial cell membrane and plasmids to pass into the cell, the pre-chilled eppendorf tube containing the mixture of competent cells and plasmids was incubated in a 42˚C water bath for 30 seconds. Then, the reaction was rapidly cooled on ice for 2 minutes. 250 µl of room temperature SOC medium (Novagen) was added into the reaction tube which was shaked at 250 rpm at 37°C for 30-60 minutes prior to plating. 50 µl and 100 µl of the transformation mixture was spread on LB media agar (Table 3.5) containing 50 µg/ml carbenicillin, 15 µg/ml tetracycline, 70 µg/ml X-gal and 80 µM IPTG. The plates were inverted and incubated overnight. The recombinant clones were screened using suitable antibiotic markers as well as blue-white colony screening.

The recombinant pET24-a plasmids were transformed into *E. coli* DH5α cells (Novagen), using the above protocol except that the LB agar contained 15 µg/ml kanamycin as the antibiotic selective maker.

3.2.6. Screening of recombinant clones by colony PCR

In order to select the correctly orientated insert clones of the white colonies of pETBlue-1 transformants, colony PCR was undertaken, using a vector-specific primer and one target gene primer. A colony from an agar plate was picked using a sterile pipette tip and transferred to a PCR tube containing 10 µl of sterile water. The reaction was heated up to 99 °C for 5 minutes to lyse the cells using the PCR machine with heated lid lock on a cap, preventing the tube from opening. It was then centrifuge at 12000 g for a minute to remove the cell debris. The supernatant was transferred to a new PCR tube as a solution of DNA template for the standard PCR. pETBlueUP-primer (TCACGACGTTGTAAAACGAC) was used together with a reverse primer of the target gene. Alternatively, pETBlueDOWN-primer (GTTAAATTGCTAACGCAGTCA) can be used together with a forward primer of the target gene. The standard PCR was carried out following the protocol A in Table 3.1. To sub-culture investigating clones for further experiments, the same colony was picked and streaked on an agar plate containing 50 µg/ml carbenicillin and 15 µg/ml tetracycline.

Table 3.5 Lysogeny broth (LB) media and LB agar recipes

	g/L
Tryptone	10
Yeast extract	5
Sodium chloride (NaCl)	5
Agar*	15

* Agar was added for preparing 1.5% w/v LB agar

3.2.7. Plasmid DNA extraction

After getting transformants, the recombinant plasmids were propagated in *E. coli* by growing up a selective single colony in 10 ml aliquots of LB media (Table 3.5) with appropriate antibiotics at 37°C, 250 rpm for overnight. The bacterial cells were harvested using a Sigma 3-16K table-top centrifuge at 5000 rpm for 5-10

minutes. After making copies of the recombinant clones in *E. coli*, the constructed DNA plasmids were extracted using the QIAprep Spin Miniprep Kit (QIAGEN) following the technical protocols provided by the manufacturer.

3.2.8. Analysis of DNA sequencing

The recombinant clones were verified by DNA sequencing with universal T7 promoter primers. The automated fluorescent DNA sequencing was carried out by external services (Genetic Core Facility, University of Sheffield and Source Bioscience). The chromatogram data were analysed using Finch TV (Geospiza, Inc.; Seattle, WA, USA).

3.3. Protein expression

After the sequence of the recombinant clones was confirmed, the correct clones were further used for producing the recombinant proteins. To produce the proteins, the recombinant genes encoding *B. pseudomallei* target proteins were expressed in *E. coli*.

3.3.1. Transformation of Tuner (DE3) pLacI

pETBlue-1 recombinant plasmids were transformed into a suitable expression host that was Tuner (DE3) pLacI (Novagen), which provides an additional source of *Lac* repressor.

1 µl of the pETBlue-1 recombinant plasmids was added to a 20 µl aliquot of thawed competent cells and was incubated on ice for 5 minutes. Then, it was heated in a 42 °C water bath for exactly 30 seconds and back on ice for 2 minutes. 80 µl of room temperature SOC medium (Novagen) was added and the mixture was incubated at 37°C while shaking at 250 rpm for 60 minutes. 5-50 µl of the cell suspension was spread on LB agar containing 50 µg/ml carbenicillin and 34 µg/ml chloramphenicol and the plates were incubated at 37°C overnight.

3.3.2. Transformation of BL21 (DE3)

pET24-a recombinant plasmids were transformed into *E. coli* BL21(DE3) competent cells (Novagen). The procedure of transformation was carried out as described in section 3.3.1 with an exception of using LB agar containing 15 µg/ml kanamycin.

3.3.3. Protein over-expression

An over expression test was initially performed in a 50 ml culture. Transformant colonies were initially inoculated in 10 ml of LB media containing the appropriate antibiotics (50 µg/ml carbenicillin and 34 µg/ml chloramphenicol were used for growing Tuner (DE3) pLacI with pETBlue-1 recombinants whereas 15 µg/ml kanamycin was used for growing BL21 (DE3) with pET24-a recombinants) and the bacteria were grown at 37°C, 250 rpm overnight after inoculation. 1% inoculum was added to 50 ml LB media containing suitable antibiotics as above and the culture was incubated at 37°C while shaking at 250 rpm until an OD_{600} of 0.6 was reached. Then, 50 µl of 1M IPTG (1 mM IPTG at final concentration) was added to an exponential phase growth culture and was incubated at the same temperature and shaking conditions for an additional 4 hours. The cell suspension was then transferred into 50 ml Falcon tubes. Bacterial cells were harvested using a Sigma 3-16K table-top centrifuge at 5000 rpm for 10-15 minutes. After draining off the liquid media, cell pastes were kept in -20°C for further analysis of protein expression.

In order to optimise the growing conditions for obtaining soluble protein production, the final concentration of IPTG added and incubation temperature after induction were varied if the initial conditions did not yield soluble protein.

To scale up the protein production, 5 ml of overnight pre-culture (1% inoculum) was added into 500 ml LB media containing suitable antibiotics. The culture was incubated at 37°C while shaking at 250 rpm until an OD_{600} of 0.6 was reached. 0.5 ml of 1 M IPTG (1 mM IPTG at final concentration) was added and the culture was incubated at 37°C (20°C for BPSS0945 production) while shaking at 250 rpm for an additional 4 hours prior to harvesting the cells. The culture was transferred into 250 ml centrifuge tubes and centrifuged at 10,000 rpm for 15-20 minutes in a Beckman Avanti J25i centrifuge. Cell pastes were collected in 50 ml Falcon tubes and kept in the -20°C freezer for subsequent purification.

3.3.4. Selenium-methionine (Se-Met) incorporated protein production

For BPSL0606, a Se-Met substituted protein was required for experimental phasing. To produce the Se-Met protein, the recombinant *E. coli* BL21 (DE3) carrying BPSL0606 in pET24-a was grown in LB media containing 15 µg/ml kanamycin at 37°C while shaking 250 rpm until OD_{600} ~ 0.6. The bacterial cells were harvested by centrifugation at 10000 rpm for 15-20 minutes at 4 °C and the liquid media were drained off. The cells were resuspended in minimal media M9 (Table 3.6) and the cell suspension was then transferred into new flasks containing 500 ml minimal media M9 with Se-Met and all other supplements plus 15 µg/ml kanamycin (Table3.6). When the OD_{600} of the bacterial culture in minimal media reached 0.6, 0.5 ml of 1 M IPTG (1 mM IPTG at final concentration) was added for induction and the culture was incubated at 37°C while shaking at 250 rpm for an additional 4 hours prior to harvesting.

Table 3.6 Minimal media preparation for Se-Met substituted protein production

M9 media	per litre
Na_2HPO_4	6.78 g
KH_2PO_4	3 g
NaCl	0.5 g
NH_4Cl	1 g
Glycerol	5 ml
Supplements	Final concentration
$CaCl_2$	0.1 mM
$MgSO_4$	2 mM
Thiamine hydrochloride	2 µg/ml
L-lysine	100 µg/ml
L-threonine	100 µg/ml
L-phenylalanine	100 µg/ml
L-isoleucine	100 µg/ml
L-valine	100 µg/ml
L-leucine	100 µg/ml
Se-Met	50 µg/ml

M9 minimal media was prepared and sterilised by autoclaving. All supplements were prepared in stock solutions, which were sterilised by a 0.2 μm pore size syringe filter and the required amounts of sterilised stock solutions were added to the autoclaved minimal media before use.

3.3.5. Protein expression analysis

In order to analyse the small-scale over-expression of the recombinant proteins, 1.5 ml of samples were taken from the culture before and after induction of IPTG. Bacterial cells were pelleted by a bench-top centrifuge at 13,000 rpm for 5 minutes. Proteins from the cell pellets were extracted and analysed by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS PAGE).

3.3.5.1. Cell disruption using Bugbuster™

Bugbuster™ (Novagen) was used to extract proteins in the cells by cell lysis. Cell pellets from 1.5 ml of bacterial culture were resuspended with a mixture of 4 μl Bugbuster™ nuclease and 100 μl Bugbuster™ protein extraction reagent. The reaction was mixed gently by pipetting the mixture and incubated on a rocking platform for 15 minutes at room temperature. The supernatant was then separated by centrifugation at 13,000 rpm for 5 minutes using a top bench centrifuge. The pellets containing the insoluble fraction of the protein were resuspended with 100 μl 50 mM Tris-HCl pH 8. The soluble fraction and the insoluble fraction were further analysed by SDS PAGE.

3.3.5.2. Cell disruption by ultra-sonication

Cell pastes were resuspended in 50 mM Tris-HCl pH 8 with a ratio of 1:10 (1 g of cell pastes : 10 ml of buffer). The cell suspension was transferred to plastic bottles on ice and the bacterial cells were disrupted by ultra-sonication on ice. The ultra-sonication was performed 3 cycles of 20 seconds each cycle with maximum amplitude and the samples were allowed to cool down between each cycle. Cell debris was separated from supernatant using in a Beckman Avanti J25i centrifuge at 24500 rpm for 10-15 minutes at 4 °C. The supernatant was subsequently analysed for protein concentration and proceeded to purification trials.

3.3.5.3. Bio-Rad protein assay

To determine the protein concentration by the Bradford method (Bradford, 1976), 1-20 µl of protein solution was added into a 1 ml plastic cuvette. 0.8 ml MilliQ water and 0.2 ml Bio-Rad reagent were added and the sample was mixed by inversion. The sample was then put into a spectrophotometer and the absorbance of the sample at 595 nm (A_{595}) was measured against the reference set to zero with the solution of MilliQ water and Bio-Rad reagent. A read out value between 0.1-0.7 was recommended for reliability. An estimated protein concentration was calculated from the following formula:

$$Protein\ concentration\ (mg/ml) = \frac{A_{595} \times 15}{the\ volume\ of\ the\ protein\ sample\ (\mu l)}$$

where 15 is an estimated coefficient based on a calibration curve using a standard protein.

3.3.5.4. Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS PAGE).

To visualise proteins for size approximation, about 20 µg protein sample was added with 10x reducing agent, 4x NuPAGE sample buffer (Invitrogen) and MilliQ water, if required, to a final volume of 20 µl. The mixture was heated in a 100 °C heating block prior to loading and running on SDS polyacrylamide gel consisting of 12% resolving gel and 6% stacking gel (Table 3.7). All samples to be analysed were loaded on to the vertical gel with an additional lane of Mark12™ unstained standard (Invitrogen) as a protein ladder. The gel was run against 1x SDS buffer (diluted from 10x stock solution, Table 3.8) at 200 Volts for 50 minutes. Then, the gel was removed from the gel cassette and rinsed with water before staining. The gel was soaked in to staining solution (0.1% (w/v) Coomassie Brilliant Blue in 20% (v/v) methanol plus 10% (v/v) acetic acid solution) until the gel colour was saturated and sufficient to visualise. Then, it was destained in destaining solution (20% (v/v) methanol plus 10% (v/v) acetic acid solution) until its background was clear.

Table 3.7 SDS polyacrylamide gel recipe

12% resolving gel	
30% Acrylamide/bisacrylamide (29:1)	2.5 ml
1M Tris-HCl buffer, pH 8.8	2.35 ml
10% SDS solution	62.5 µl
10% Ammonium persulphate solution	62.5 µl
TEMED	6.25 µl
MilliQ water	1.28 ml
6% stacking gel	
30% Acrylamide/bisacrylamide (29:1)	0.75 ml
1M Tris-HCl buffer, pH 6.8	0.47 ml
10% SDS solution	37.5 µl
10% Ammonium persulphate solution	37.5 µl
TEMED	3.75 µl
MilliQ water	2.46 ml

Table 3.8 10x SDS buffer recipe

	Per liter
Glycine	144 g
Tris	30g
SDS	10g

3.4. Protein purification, crystallisation and data collection

For each of the proteins studied, purification was undertaken using a combination of ion exchange chromatography and gel filtration chromatography. Crystals were grown using commercial screens (QIAGEN) using a Hydra II Plus One robot. Data collection was collected using both an in-house Rigaku MicroMax 007 micro-focus copper rotating anode generator and Diamond synchrotron radiation.

Details for each protein are given in the following chapters; section 4.4.4 (BPSS0945), section 4.6.3.3, 5.1 and 5.2 (BPSL0606).

3.5. Data processing

All X-ray diffraction images were autoindexed and integrated by Mosflm (Leslie, 2006). The diffraction data were scaled and merged by SCALA (Evans, 2006) . At the Diamond beamlines, diffraction data were automatically processed using XDS in the xia2 pipeline (Winter, 2010). The Matthews_coeff program in the CCP4 suite (Collaborative Computational Project, 1994) was used to analyse the unit cell contents on a basis of the unit cell volume and protein size (Matthews, 1968).

3.6. Structure determination

The software suite Phenix (Adams *et al*, 2010) was used to determine the protein structure using molecular replacement and the MAD method. The best structure solution was obtained and the protein chains were re-built using the CCP4 suite (Collaborative Computational Project, 1994) and Coot (Emsley *et al*, 2010). REFMAC5 was used for structure refinement (Murshudov *et al*, 2011). Structure validation was carried out using RAMPAGE in the CCP4 suite (Lovell *et al*, 2003).

3.7. Ligand observation by nuclear magnetic resonance (NMR)

Freshly purified BPSL0606 protein was prepared in 20 mM Na-phosphate buffer pH 6.5 plus 0.1 M NaCl for ^{1}H-NMR experiment. The concentration was calculated using the measured absorbance at 280 nm divided by an extinction coefficient predicted from ProtParam tool. Acetyl coenzyme A solution was prepared at the same molar concentration of the protein. A mixture of purified protein and acetyl coenzyme A was prepared in a 1:1 molar ratio. All samples were loaded into NMR tubes and added with 50 µl heavy water (D_2O) prior to loading the samples to a Bruker Avance 600MHz NMR spectrometer equipped with a 5mm TXI 13C/15N Z-gradient probe. ^{1}H NMR spectra of samples were observed.

Chapter 4: Target Selection, Cloning, Protein Expression and Purification

This chapter describes target selection, gene amplification, cloning of target genes, protein over-expression and protein purification. Although the desired outcome of this project would be a molecular structure of each target protein, it is unrealistic to assume that this will occur, with failures possible at any experimental step. Thus, the gene amplification and cloning of BPSL3022, BPSS0945, BPSS0603, BPSS0683, BPSL0606, BPSS0238 and BPSS1416 are described. The protein expression of BPSL3022, BPSS0945, BPSS0603, BPSS0683 and BPSL0606 are detailed and the protein purification of BPSS0945 and BPSL0606 are discussed.

4.1. Target selection

A list of target genes in this study was created from a selection of potential drug targets, which are detailed below.

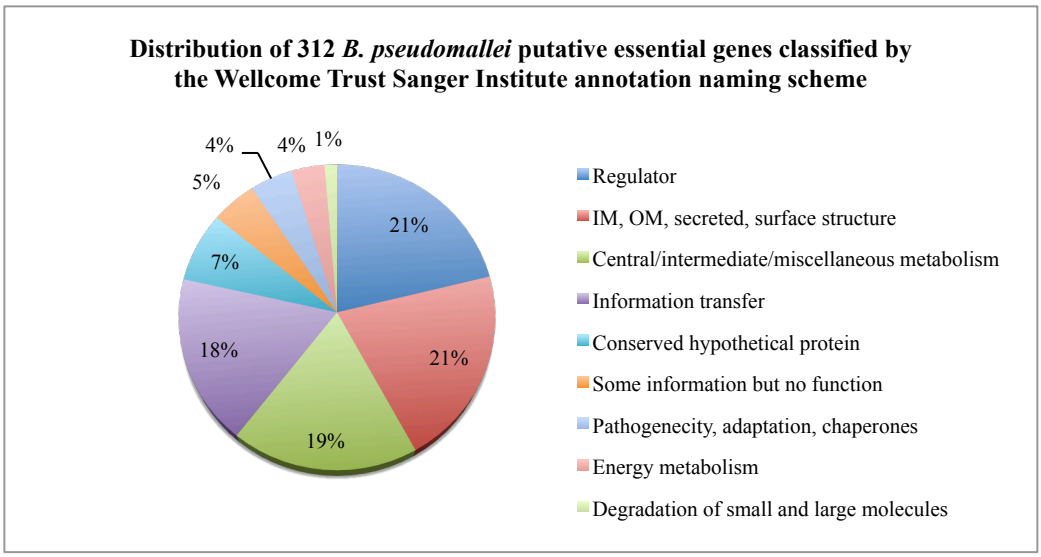
Su *et al* (2008) have experimentally identified 109 immunogenic proteins by western blotting a *B. pseudomallei* D286 genomic expression library against the pooled sera of melioidosis patients, as a primary antibody screening. The proteins that reacted with the serum of the melioidosis patients may thus be involved in the pathogenesis and virulence of a *B. pseudomallei* infection. Based on a database of conserved domains (Marchler-Bauer *et al*, 2011), these immunogenic proteins have been classified into different functional groups, including cell envelope biogenesis (13%), cell mobility and secretion (6%), transcription (6%), amino acid transport and metabolism (11%), inorganic ion transport and metabolism (3%), protein metabolism (8%), energy production and conversion (8%), nucleic acid metabolism and repair (5%), miscellaneous metabolic proteins (16%), other transporters (7%) and uncharacterised hypothetical proteins (17%).

In conjunction, 312 putative essential genes of *B. pseudomallei* have been identified computationally (Chong *et al*, 2006). The *B. pseudomallei* genome was compared with the human genome sequence to identify those *B. pseudomallei* genes that do not have a homologue in humans. These genes were further compared with known essential gene sequences in the database of bacterial essential genes (Zhang *et al*, 2004) to obtain a list of putative essential genes in *B. pseudomallei*. These genes were classified into different functional classes according to the Wellcome Trust Sanger Institute annotation naming strategies, including regulator proteins (21.2%), inner membrane/outer membrane/secreted/ surface structure (20.8%), information transfer (transcription/translation/DNA/ RNA modification) (17.9%), central/intermediate/miscellaneous metabolism (18.6%), energy metabolism (3.5%), degradation of small and large molecules (1.3%), pathogenicity/adaptation/ chaperones (4.5%), some information but no function (4.8%) and conserved hypothetical proteins (7.4%).

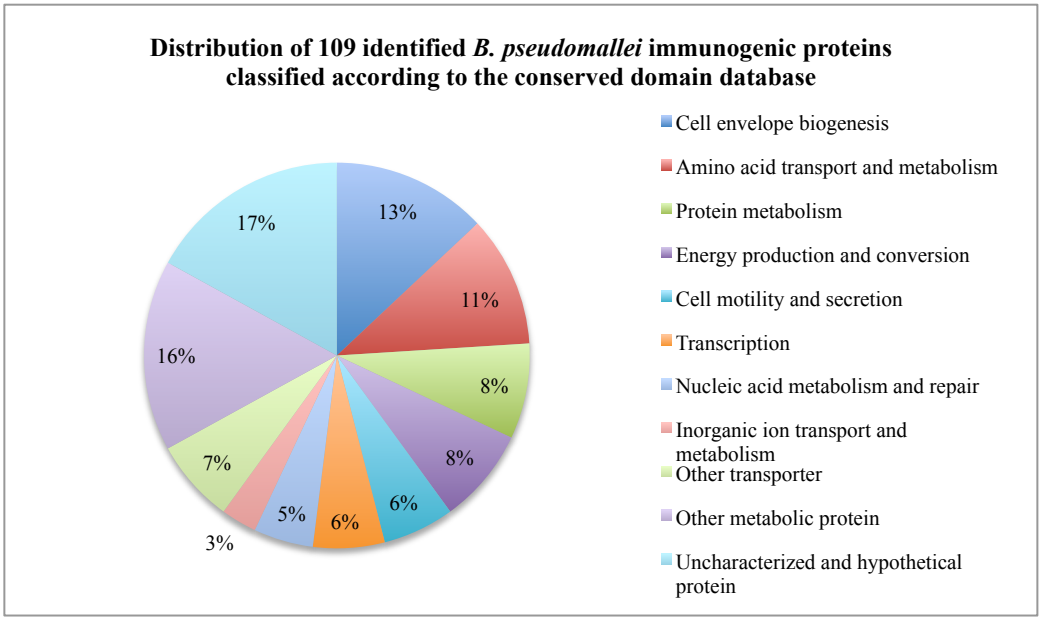
To make a list of a potential drug candidates in this study, the 109 immunogenic proteins from the study of Su et al (2008) were compared with the 312 putative essential genes of the study of Chong *et al* (2006) using the *B. pseudomallei* locus tag identity (BPS_XXXX). The intersection between these two identified groups gave a list of four proteins that elicit an immune response in infected individuals, and are also predicted to be essential for *B. pseudomallei* survival (Figure 4.1). These proteins (BPSL3022, BPSS0945, BPSS0238 and BPSS1416), which may play a role in pathogenesis or virulence and are also attractive drug targets, were thus selected as the first four proteins to be characterised (Table 4.1)

A further three uncharacterised proteins, BPSS0603, BPSS0683 and BPSL0606, which were also immunogenic proteins but were not putative essential genes, were selected and added to the list after failing to obtain crystals from any of the first four proteins for crystallographic studies.

a)


Distribution of 312 B. pseudomallei putative essential genes classified by the Wellcome Trust Sanger Institute annotation naming scheme
21%
21%
19%
18%
7%
5%
4%
4%
1%
Regulator
IM, OM, secreted, surface structure
Central/intermediate/miscellaneous metabolism
Information transfer
Conserved hypothetical protein
Some information but no function
Pathogenecity, adaptation, chaperones
Energy metabolism
Degradation of small and large molecules

b)


Distribution of 109 identified B. pseudomallei immunogenic proteins classified according to the conserved domain database
13%
11%
8%
8%
6%
6%
5%
3%
7%
16%
17%
Cell envelope biogenesis
Amino acid transport and metabolism
Protein metabolism
Energy production and conversion
Cell motility and secretion
Transcription
Nucleic acid metabolism and repair
Inorganic ion transport and metabolism
Other transporter
Other metabolic protein
Uncharacterized and hypothetical protein

c)

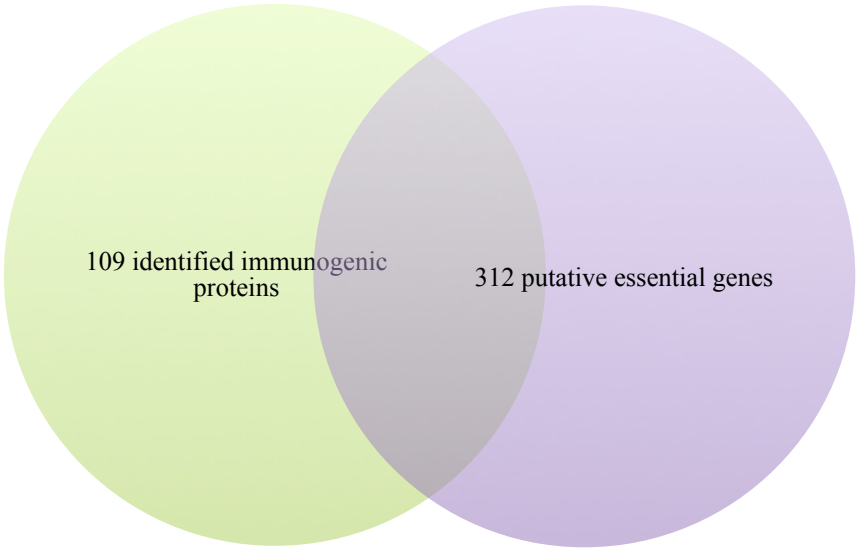


Figure 4.1 An initial target selection:

a) 312 putative essential genes identified for potential drug target were grouped by their relative functions (Chong *et al*, 2006).

b) The experimentally identified 109 immunogenic proteins were classified into their conserved domain related groups (Su *et al*, 2008).

c) A Venn diagram showed the intersection between these two identified targets, which includes BPSL3022, BPSS0945, BPSS0238 and BPSS1416.

Table 4.1 A list of selected targets

Locus tag	Annotation	Gene size (base pair)	No. of Amino acid	Molecular weight (kDa)
BPSL3022	Cell division protein FtsQ	753	250	27.93
BPSS0945	Subfamily M23B unassigned peptidase	948	315	32.03
BPSS0238	Penicillin binding protein	2163	720	76.75
BPSS1416	Hypothetical protein	1404	467	47.72
BPSS0603	Hypothetical protein	303	100	10.99
BPSS0683	Hypothetical protein	357	118	12.99
BPSL0606	Hypothetical protein	651	216	23.49

4.2. Gene amplification

Oligonucleotide primers for target gene amplification were designed according to their open reading frames of their individual nucleotide sequence based on the completely sequenced genome of *B. pseudomallei* strain K96423.

4.2.1. Primer design

PCR primers were designed compatible to vectors used for subsequent cloning. Two different cloning methods were used in this study.

The first method is TA cloning using the pETBlue-1 vector, which facilitates cloning and enables high-level recombinant protein expression in *E.coli* (Novagen). This system requires the insert to start with an ATG start codon so that translation initiation occurs, with an optimal spacing of 8 base pairs between the ribosome binding site and the start codon. Therefore, the forward primers were designed to start with ATG and the reverse primers were designed complementary to their own sequence. This method was used for cloning of BPSL3022, BPSS0945, BPSS0238, BPSS1416, BPSS0603 and BPSS0683.

The second method of cloning used in this project, was restriction enzyme cloning, with pET24-a vector, which is a cloning/expression vector that provides a C-terminal His tag fusion (Novagen). For the restriction enzyme cloning, primers were designed with the additional recognition sequence of restriction enzymes used in the cloning. The target sequences were firstly analysed using NEB cutter (Vincze *et al*, 2003) to identify restriction sites present in the target sequences. The primers were then designed to contain restriction sites not present in the target gene. The forward primer of BPSL3022, BPSS0238, BPSS1416 and BPSL0606 contained *Nde*I (CATATG). The reverse primers for different constructs of BPSL3022 either had *Eco*RI (GAATTC) or *Xho*I (CTCGAG) sites, whereas the reverse primers for BPSS0238, BPSS1416 and BPSL0606 contained *Bam*HI (GGATCC). A list of primers used in this study is shown in Table 4.2.

Table 4.2 A list of primers used in this work (underline-restriction site)

Locus tag Annotation	Cloning method	5'-3' Oligonucleotides	
		Forward primer	Reverse primer
BPSL3022 Cell division protein FtsQ			
BPSL3022_1	TA	ATGCGGGAAATCCGCATCG 19 bp Tm 60 °C, %GC 57.9	GCTTACATGCCCGCCGC 17 bp Tm 58 °C, %GC 70.6
BPSL3022_2	TA		TTACTTCTTGCCCTTGTCGGT 21 bp Tm 62 °C, %GC 47.6
BPSL3022_1	RE	GCGTTCCATATGCGGGAAATC 21 bp Tm 64 °C, %GC 52.4	GGTGAATTCTTACATGCCCGC 21 bp Tm 64 °C, %GC 52.4
BPSL3022_2	RE		TGATGAATTCTTACTTCTTGCCCT 24 bp Tm 64 °C, %GC 37.5
BPSL3022_3	RE		GTCCTCGAGCTTCTTGCCCT 20 bp Tm 64 °C, %GC 60
BPSS0945 Subfamily M23B unassigned peptidase	TA	ATGAGCAAGAGCGAGATCGA 20 bp Tm 60 °C, %GC 50	CTAGCCCTGCTGCCGGC 17 bp Tm 60 °C, %GC 76.5

Table 4.2 A list of primers used in this work (underline-restriction site) (continued)

Locus tag Annotation	Cloning method	5'-3' Oligonucleotides	
		Forward primer	Reverse primer
BPSS0238 Penicillin binding protein	TA	ATGCCAACGAAGGATTTCGAC 21 bp Tm 62 °C, %GC 47.6	TCAGTCTTCGGTGCGAAACA 20 bp Tm 60 °C, %GC 50
	RE	GTACGCCGCATATGCCAAC 19 bp Tm 60 °C, %GC 57.9	GGATCCCCGCTCAGTCTTC 19 bp Tm 62 °C, %GC 63.2
BPSS1416 Hypothetical protein	TA	ATGCCTGCCTCCATGAGC 18 bp Tm 58 °C, %GC 61.1	TCATGCGGCGCCCTCCA 17 bp Tm 58 °C, %GC 70.6
	RE	CCGCATATGCCTGCCTCCAT 20 bp Tm 64 °C, %GC 60	AGGATCCTCATGCGGCGCCC 20 bp Tm 68 °C, %GC 70
BPSS0603 Hypothetical protein	TA	ATGATAATGGAGGCAAGCATGT 22 bp Tm 62 °C, %GC 40.9	GTTACGTGCTCGCCGC 16 bp Tm 54 °C, %GC 68.8
BPSS0683 Hypothetical protein	TA	ATGCGCCTCACCATTCGAAT 20 bp Tm 60 °C, %GC 50	TCAGCGCACGATGGTGAAT 19 bp Tm 58 °C, %GC 52.6
BPSL0606 Hypothetical protein	RE	GAATCTCCCATATGGACTGG 20 bp Tm 60 °C, %GC 50	GGATCCTCATCGGTTCATTC 20 bp Tm 60 °C, %GC 50

NB: TA = TA cloning
RE = Restriction enzyme cloning

4.2.2. Polymerase chain reaction (PCR)

The amplification of these target genes was carried out using the polymerase chain reaction (PCR). The genomic DNA of *B. pseudomallei* strain D286, kindly provided by the Universiti Kebangsaan Malaysia, was used as a template. The optimization of PCR was carried out for individual gene amplification by varying the annealing temperature and the percentage of dimethyl sulfoxide (DMSO). As the *B. pseudomallei* genome contains high GC levels (68%), the addition of DMSO was used in the PCR in order to facilitate base pairing disruption between double stranded DNA with high GC contents (Kang *et al*, 2005; Musso *et al*, 2006). The PCR conditions for each gene are shown in Table 4.3.

Table 4.3 PCR conditions for target gene amplification

	Annealing temperature (°C)	DMSO (% V/V)
BPSL3022_1/TA	56	-
BPSL3022_2/TA	56	-
BPSL3022_1/RE	60	-
BPSL3022_2/RE	56	5
BPSL3022_3	58	2
BPSS0945	56	5
BPSS0238	56	10
BPSS1416	56	5
BPSS0603	53	-
BPSS0683	60	-
BPSL0606	56	2

4.3. BPSL3022-Cell division protein FtsQ

BPSL3022 truncation: periplasmic protein constructs

In the *B. pseudomallei* genome, BPSL3022 has been annotated as the cell division protein FtsQ (Winsor *et al*, 2008). The *ftsQ* gene in *E. coli* is found in a cell division gene cluster that is involved in septum formation and synthesis of cell wall components. The FtsQ protein (31 kDa) consists of a 24 N-terminal amino acid residue cytoplasmic tail, residues 25-49 embedded in the cytoplasmic membrane and the remainder forming a C-terminal periplasmic domain of 226 residues (Carson *et al*, 1991). The structures of FtsQ from *E. coli* and *Yersinia enterocolitica* have been successfully determined (Van den Ent *et al*, 2008). Both these proteins are composed of two domains. The topology of the α-domain close to the membrane is β-α-α-β-β and the C-terminal β-domain consists of nine β-strands with two long α-helices (Figure 4.2). The constructs for expressing the proteins, that crystallised into these two structures of the periplasmic domain of FtsQ in *E. coli* and *Y. enterocolitica*, contain residues 58-276 and residues 59-285 for *E. coli* and *Y. enterocolitica* FtsQ, respectively, each lacking the N-terminal hydrophobic region.

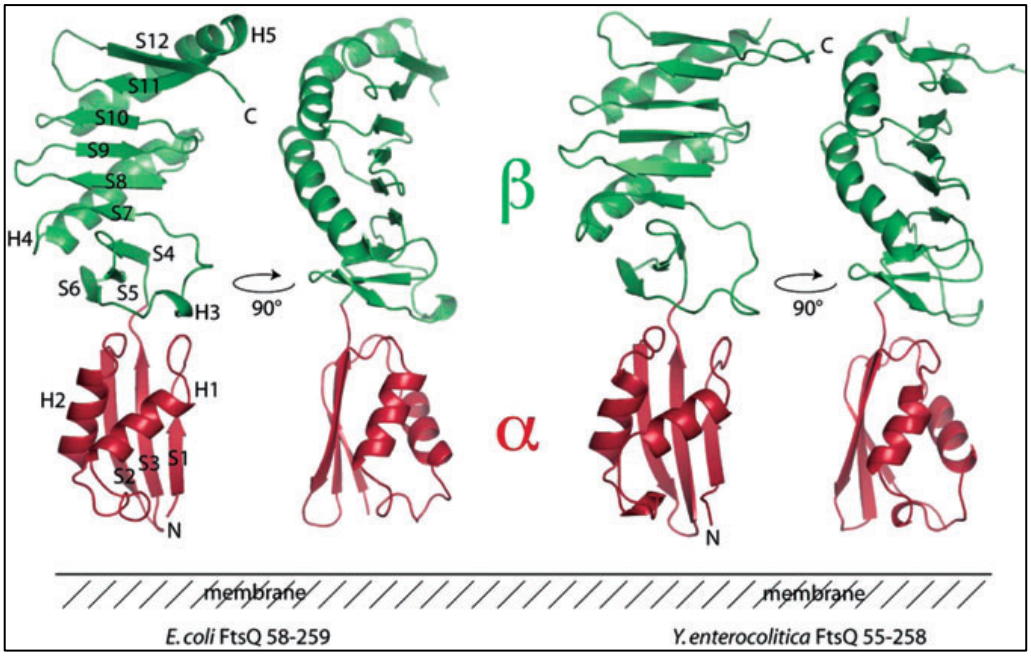


Figure 4.2 Crystal structures of periplasmic FtsQ in *E. coli* and *Y. entericolitica*: *E.coli* and *Y.entericolitica* FtsQ have 56% sequence identity and the rmsd of their superposed structures (Cα) is 1.63 Å. This figure is taken from Van den Ent *et al* (2008).

The protein sequence of FtsQ in *B. pseudomallei* (BPSL3022) was analysed using hydropathy plot, which showed that the protein contains a N-terminal hydrophobic part, which is likely to be a membrane anchor (Figure 4.3). Kyte-Doolittle scale plots show the distribution of hydrophobic and hydrophilic residues on the protein sequence and positive values represent the hydrophobic character of the protein. A score of more than 1.6 indicates a potential hydrophobic membrane-spanning domain. This corresponds to a prediction of transmembrane helices, based on a hidden Markov model by the TMHMM server v.2.0 (Krogh *et al*, 2001). The TMHMM annotation probabilities for the sequence of BPSL3022 are shown in Figure 4.4.

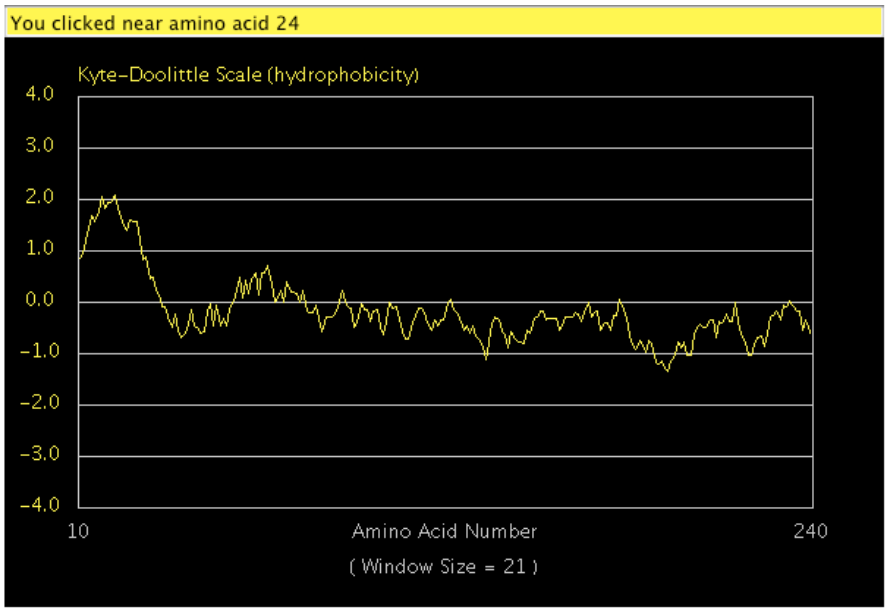


Figure 4.3 A hydropathy plot for BPSL3022, showing the N-terminal hydrophobic region (residues 1-24) of this protein. This figure was produced using the hydropathy plot server (http://www.vivo.colostate.edu/molkit/hydropathy/index.html).

```
# Sequence Length: 250
# Sequence Number of predicted TMHs:  1
# Sequence Exp number of AAs in TMHs: 22.31402
# Sequence Exp number, first 60 AAs:  22.30853
# Sequence Total prob of N-in:        0.43501
# Sequence POSSIBLE N-term signal sequence
Sequence        TMHMM2.0        outside      1     9
Sequence        TMHMM2.0        TMhelix     10    32
Sequence        TMHMM2.0        inside      33   250
```

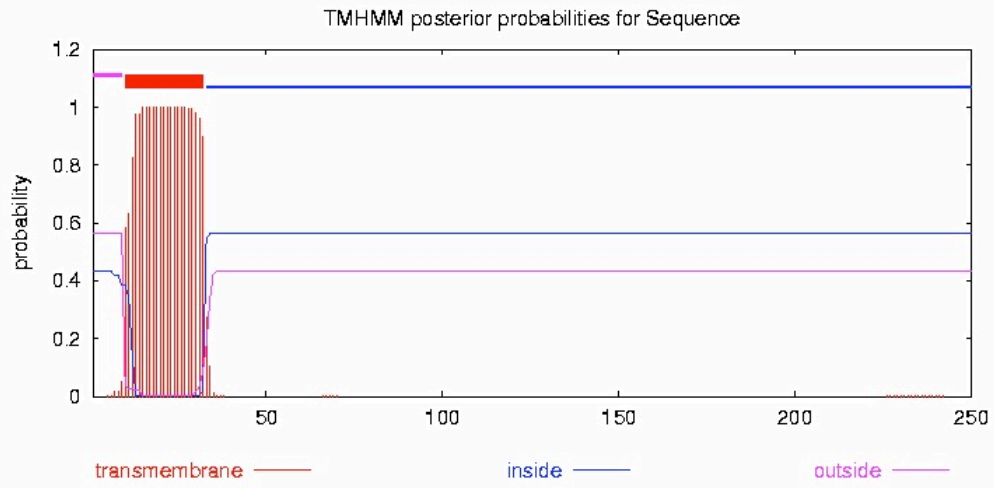


Figure 4.4 Transmembrane prediction of BPSL3022: Residues 10 to 32 of BPSL3022 are likely to form a transmembrane helix. This figure was produced using TMHMM server (Krogh *et al*, 2001).

When the secondary structure of BPSL3022 was predicted using mGenTHREADER PREDICTION (McGuffin *et al*, 2000), the top hits had approximately 30% identity to the crystal structures of *E. coli* and *Y. enterocolitica*. Thus, the protein sequence of BPSL3022 was aligned with the sequences of the crystal structures of periplasmic FtsQ in *E. coli* and *Y. enterocolitica* using ClustalW (Thompson *et al*, 1994), as shown in Figure 4.5.

Using this alignment, two constructs of BPSL3022 were designed, in an attempt to optimise the probability of crystallisation. According to the BPSL3022 sequence, one truncated form of BPSL3022 (named BPSL3022_1) was constructed from residue 40 (arginine, R) to residue 239 (methionine, M), where the sequence similarity to the FtsQ homologue structures ends, and the other construct (named BPSL3022_2) started from residue 40 to the C-terminus, residue 250 (lysine, K) (Figure 4.5). Therefore, the oligonucleotide primers of the truncated BPSL0606 were designed corresponding to the open reading frame from the residue 40 with additional start codon ATG to amplify these gene fragments.

```
Y._enterocoliti ----------------------------------------SKLVVTGERHYTTNDDIRQAI
E._coli_gi_1460 ----------------------------------------SKLVLTGERHYTRNDDIRQSI
B._pseudomallei MWNNVRQLNLAASALYALLLLVLAAAGCYWLIQRPAFALREIRIDGDTEHINAPTVRAGV
consensus       ----------------------------------------sklvvtGerhyt-nddiRqai
                1........10........20........30........40........50.........

Y._enterocoliti LSLGAPGTFMTQDVNIIQQQIER-LPWIQQASVRKQWPDELKIHLVEYVPFARWNDLHMV
E._coli_gi_1460 LALGEPGTFMTQDVNIIQTQIEQRLPWIKQVSVRKQWPDELKIHLVEYVPIARWNDQHMV
B._pseudomallei VGR-LKGNFFTVDLDLARVAFEQ-MPWVRHASVRRVWPNALAVTLEEYKPLGTWGNDQLV
consensus       lalg-pGtFmTqDvniiq-qiEq-lPWikqaSVRkqWPdeLkihLvEYvPiarWnd-hmV
                61.......70........80........90........100.......110........

Y._enterocoliti DEQGRSFSVPSERVGKQKLPLLYGPEGSEQDVLEGYRAINKVLAANKYQLKMVAMSARHS
E._coli_gi_1460 DAEGNTFSVPPERTSKQVLPMLYGPEGSANEVLQGYREMGQMLAKDRFTLKEAAMTARRS
B._pseudomallei SVDGELFTAN-QGELDAELPSFDGPEGSAKEVVARYRDFAKWFAPIHATPEEVTLSPRYA
consensus       d-eG-sFsvp-er--kq-LPllyGPEGSaqeVl-gYReigkvlA--kytlkevamsaR-s
                121......130.......140.......150.......160.......170........

Y._enterocoliti WQLALDNDVRLELGR----DDRMGRLQRFIELYPMLQQQPD---KRVSYVDLRYETGAAI
E._coli_gi_1460 WQLTLNNDIKLNLGR----GDTMKRLARFVELYPVLQQQAQTDGKRISYVDLRYDSGAAV
B._pseudomallei WTVKLSNGMQVELGRERNSDTLPDRIQRLVAAWPSVTQRWG---GDIEYADLRYPNGFAI
consensus       Wql-L-NdvrleLGR----dd-m-RlqRfvelyPmlqQq-----krisYvDLRYetGaAi
                181......190.......200.......210.......220.......230........

Y._enterocoliti GWAPVFIGSQGGEPPVNGQQNSNPQQNQAQAKQQ
E._coli_gi_1460 GWAPL--------PPEESTQ----QQNQAQAEQQ
B._pseudomallei RAAGMR-----------------FLTDTDKGKK-
consensus       gwApv--------pp----q----qqnqaqakqq
                241......250.......260.......270..
```

Figure 4.5 Protein sequence alignment of the crystal structures of the periplasmic FtsQ protein from *E. coli* (second line) and *Y. enterocolitica* (first line) and FtsQ from *B. pseudomallei* (third line); The constructs that crystallised for FtsQ in *E. coli* and *Y. enterocolitica* were considered in order to design a construct of periplasmic FtsQ in *B. pseudomallei*, on the basis of sequence similarity. Completely conserved residues are shown in green, identical residues are in yellow and similar residues are shown in cyan. This figure was created using Boxshade (Hofmann & Baron).

The protein sequences of the two BPSL3022 truncates were analysed by ProtParam (ExPASy server). The short truncated FtsQ consists of 201 amino acids with a molecular weight of 22.6 kDa and a theoretical pI of 5.3, whereas the long truncated FtsQ is 212 amino acids long and has a molecular weight of 23.9 kDa with a theoretical pI of 5.7 (Figure 4.6).

```
MREIRIDGDT EHINAPTVRA GVVGRLKGNF FTVDLDLARV AFEQMPWVRH ASVRRVWPNA
LAVTLEEYKP LGTWGNDQLV SVDGELFTAN QGELDAELPS FDGPEGSAKE VVARYRDFAK
WFAPIHATPE EVTLSPRYAW TVKLSNGMQV ELGRERNSDT LPDRIQRLVA AWPSVTQRWG
GDIEYADLRY PNGFAIRAAG M
```

Number of amino acids: 201 **Molecular weight:** 22584.4 **Theoretical pI:** 5.30

Amino acid composition:

Ala (A)	21	10.4%	Ile (I)	7	3.5%	Tyr (Y)	5	2.5%
Arg (R)	18	9.0%	Leu (L)	16	8.0%	Val (V)	18	9.0%
Asn (N)	8	4.0%	Lys (K)	5	2.5%	Pyl (O)	0	0.0%
Asp (D)	13	6.5%	Met (M)	4	2.0%	Sec (U)	0	0.0%
Cys (C)	0	0.0%	Phe (F)	8	4.0%			
Gln (Q)	6	3.0%	Pro (P)	12	6.0%	(B)	0	0.0%
Glu (E)	15	7.5%	Ser (S)	8	4.0%	(Z)	0	0.0%
Gly (G)	16	8.0%	Thr (T)	11	5.5%	(X)	0	0.0%
His (H)	3	1.5%	Trp (W)	7	3.5%			

Total number of negatively charged residues (Asp + Glu): 28
Total number of positively charged residues (Arg + Lys): 23

Atomic composition:

Carbon	C	1010
Hydrogen	H	1557
Nitrogen	N	287
Oxygen	O	296
Sulfur	S	4

Formula: $C_{1010}H_{1557}N_{287}O_{296}S_4$
Total number of atoms: 3154

Extinction coefficients:
Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.
Ext. coefficient 45950
Abs 0.1% (=1 g/l) 2.035, assuming ALL Cys residues appear as half cystines

a)

```
MREIRIDGDT EHINAPTVRA GVVGRLKGNF FTVDLDLARV AFEQMPWVRH ASVRRVWPNA
LAVTLEEYKP LGTWGNDQLV SVDGELFTAN QGELDAELPS FDGPEGSAKE VVARYRDFAK
WFAPIHATPE EVTLSPRYAW TVKLSNGMQV ELGRERNSDT LPDRIQRLVA AWPSVTQRWG
GDIEYADLRY PNGFAIRAAG MRFLTDTDKG KK
```

Number of amino acids: 212 **Molecular weight:** 23874.9 **Theoretical pI:** 5.71

Amino acid composition:

Ala (A)	21	9.9%	Ile (I)	7	3.3%	Tyr (Y)	5	2.4%
Arg (R)	19	9.0%	Leu (L)	17	8.0%	Val (V)	18	8.5%
Asn (N)	8	3.8%	Lys (K)	8	3.8%	Pyl (O)	0	0.0%
Asp (D)	15	7.1%	Met (M)	4	1.9%	Sec (U)	0	0.0%
Cys (C)	0	0.0%	Phe (F)	9	4.2%			
Gln (Q)	6	2.8%	Pro (P)	12	5.7%	(B)	0	0.0%
Glu (E)	15	7.1%	Ser (S)	8	3.8%	(Z)	0	0.0%
Gly (G)	17	8.0%	Thr (T)	13	6.1%	(X)	0	0.0%
His (H)	3	1.4%	Trp (W)	7	3.3%			

Total number of negatively charged residues (Asp + Glu): 30
Total number of positively charged residues (Arg + Lys): 27

Atomic composition:

Carbon	C	1067
Hydrogen	H	1652
Nitrogen	N	304
Oxygen	O	313
Sulfur	S	4

Formula: $C_{1067}H_{1652}N_{304}O_{313}S_4$
Total number of atoms: 3340

Extinction coefficients:
Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.
Ext. coefficient 45950
Abs 0.1% (=1 g/l) 1.925, assuming ALL Cys residues appear as half cystines

b)

Figure 4.6 Protein sequence analysis of BPSL3022 constructs by ProtParam: a) Characteristics of the short truncated BPSL3022 (residues 40-239); b) Characteristics of the long truncated BPSL3022 (residues 40-250)

4.3.1. Gene amplification

The target genes encoding the periplasmic part of *B. pseudomallei* FtsQ were amplified from the genomic DNA of *B. pseudomallei* strain D286 using the oligonucleotide primers designed according to the open reading frame that encodes from residue 40 of the full-length protein with additional start codon (ATG) to fulfill the requirement of pETBlue-1 vector system. The forward primer (5'-ATGCGGGAAATCCGCATCG-3') and reverse primer (5'- GCTTACATG CCCGCCGC-3') were used in a standard PCR to obtain the short truncated FtsQ, while the same forward primer was used together with the reverse primer (5'-TTACTTCTTGCCCTTGTCGGT-3') for obtaining the long truncated construct. The PCR products for both gene fragments amplification were achieved using an annealing temperature of 56 °C with an expected size of approximately 600 base pairs (Figure 4.7).

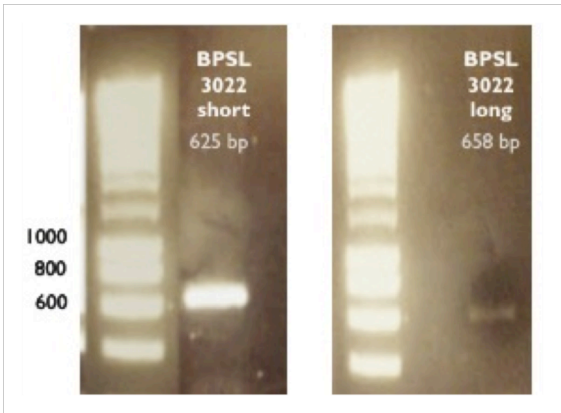


Figure 4.7. PCR products of truncated BPSL3022 gene amplification on a 1% agarose gel: The expected size of the PCR products obtained was in the region of 600 base pairs.

4.3.2. Cloning of BPSL3022_1 and BPSL3022_2

The PCR products of BPSL3022_1 and BPSL3022_2 were cloned into the pETBlue-1 vector by joining the given 3'-dU overhang of the linearised pETBlue-1™ AccepTor vector and the 3'-dA overhangs of the PCR products. The linearised pETBlue-1™ AccepTor vector provides single 3'-dU overhangs that can be ligated with an insert with 3'-dA overhangs, generated during PCR with a DNA polymerase lacking 3'->5'exonuclease activity. After ligation, the recombinant plasmids were transformed into *E.coli* NovaBlue, which enables blue/white screening. White colonies obtained are the bacteria carrying the

recombinant plasmids, as the insert interferes with a *lacZ* α-peptide open reading frame in the vector, resulting in the disruption of β-galactosidase production, which disables the cleavage of the chromogenic substrate, X-gal, on the agar plate.

Recombinant clones of BPSL3022 were selected by the blue white colony method. The recombinant plasmids can carry the insert either correctly oriented or the wrong way round. To determine the correct orientation of the insert, the white colonies were further screened by colony PCR. The use of a vector specific primer together with an insert primer for amplifying the investigating plasmid DNA yields the expected sized PCR product, if the recombinant has the correct orientation. Recombinant plasmids with a positive PCR result were propagated in *E. coli*. The recombinant plasmids were extracted and the sequence confirmed by DNA sequencing with universal T7 primers.

The sequencing results of the long truncated (BPSL3022_2) clone showed that there were two nucleotides that were different from the reference genome of *B. pseudomallei* strain K96243, resulting in one amino acid alteration at residue 79 (A to V) (Figure 4.8). On the other hand, the BPSL3022_1 short truncated construct could not be successfully cloned, since the plasmid sequencing indicated that these clones contained the long truncated insert. This may be due to an inappropriate primer design.

```
BPSL3022_24          1 ------------------------------------MREIRIDGDTEHINAPTVRAGV
gi_53720632_ref      1 MWNNVRQLNLAASALYALLLLVLAAAGCYWLIQRPAFALREIRIDGDTEHINAPTVRAGV
consensus            1 mwnnvrqlnlaasalyallllvlaaagcywliqrpafamREIRIDGDTEHINAPTVRAGV

BPSL3022_24         23 VGRLKGNFFTVDLDLARVVFEQMPWVRHASVRRVWPNALAVTLEEYKPLGTWGNDQLVSV
gi_53720632_ref     61 VGRLKGNFFTVDLDLARVAFEQMPWVRHASVRRVWPNALAVTLEEYKPLGTWGNDQLVSV
consensus           61 VGRLKGNFFTVDLDLARV-FEQMPWVRHASVRRVWPNALAVTLEEYKPLGTWGNDQLVSV

BPSL3022_24         83 DGELFTANQGELDAELPSFDGPEGSAKEVVARYRDFAKWFAPIHATPEEVTLSPRYAWTV
gi_53720632_ref    121 DGELFTANQGELDAELPSFDGPEGSAKEVVARYRDFAKWFAPIHATPEEVTLSPRYAWTV
consensus          121 DGELFTANQGELDAELPSFDGPEGSAKEVVARYRDFAKWFAPIHATPEEVTLSPRYAWTV

BPSL3022_24        143 KLSNGMQVELGRERNSDTLPDRIQRLVAAWPSVTQRWGGDIEYADLRYPNGFAIRAAGMR
gi_53720632_ref    181 KLSNGMQVELGRERNSDTLPDRIQRLVAAWPSVTQRWGGDIEYADLRYPNGFAIRAAGMR
consensus          181 KLSNGMQVELGRERNSDTLPDRIQRLVAAWPSVTQRWGGDIEYADLRYPNGFAIRAAGMR

BPSL3022_24        203 FLTDTDKGKK
gi_53720632_ref    241 FLTDTDKGKK
consensus          241 FLTDTDKGKK
```

Figure 4.8 DNA sequencing of BPSL3022_2 in the pETBlue-1 clone: The translated protein sequence of the recombinant truncated BPSL3022 (first line) was aligned with the full length protein sequence of FtsQ in *B. pseudomallei* strain K96243 (second line), showing one amino acid alteration highlighted in red (Ala to Val). This figure was created using Boxshade (Hofmann & Baron).

4.3.3. Protein expression

Expression of the truncated BPSL3022_2 in pETBlue-1 was carried out in *E. coli* Tuner™ (DE3), with 1mM IPTG induction at 37°C for 4 hours after growing to mid log phase (OD_{600} ~ 0.6). After harvesting and disrupting the bacterial cells, the protein expression profiles were analysed by SDS PAGE. The expected molecular weight of this construct is 23.9 kDa (ProtParam, ExPASy server). There was an intense band between 21.5 and 31 kDa on the SDS PAGE gel (Figure 4.9). In the host expression system, *E. coli* Tuner™ also produces chloramphenicol acetyl transferase (CAT) which is 25.7 kDa in size. Therefore, it was uncertain whether the band visualised on the gel was the target protein or the CAT protein. It was quite difficult to separate these two proteins with their similar pI (BPSL3022_2 pI 5.7, CAT pI 5.9) as there was no affinity tag on the target protein.

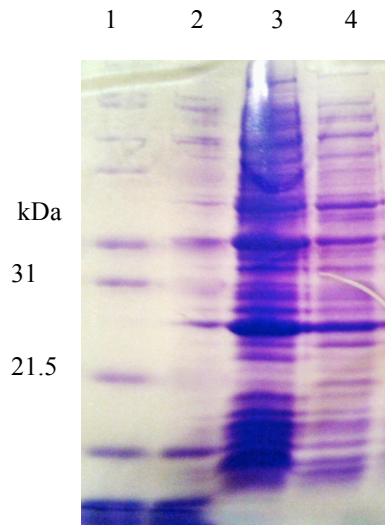


Figure 4.9 SDS PAGE BPSL3022_2 in pETBlue-1 expressed in *E. coli* Tuner (DE3) pLacI cells: lane 1 Mark12™; lane 2 pre-induction; lane 3 soluble fraction at 37°C induction for 4 hours; lane 4 insoluble fraction at 37°C induction for 4 hours

As overexpression of the CAT protein often occurs in the pETBlue-1 expression system, a different construct, BPSL3022_3, was thus designed in order to clone into the pET24-a vector system using the restriction enzyme strategy. This construct provides a C-terminal His-tag to the protein in order to ease protein purification. The forward primer (5'-GCGTTCCATATGCGGGAAATC-3') and the reverse primer (5'- GTCCTCGAGCTTCTTGCCCT-3') were used to amplify the truncated periplasmic FtsQ of *B. pseudomallei* (residue 40-250). The PCR reaction for this truncated BPSL3022 required the addition of 2% DMSO at an annealing temperature of 58°C. The expected size of the PCR product was approximately 600 base pairs on a 1% agarose gel.

The cloning of the BPSL3022_3 construct using restriction enzymes was carried out by ligating the compatible overhangs of the digested PCR product and a linearised pET24-a vector. The PCR product was digested with the restriction enzymes, *Nde*I and *Xho*I to provide the overhangs 5'…-CA^TATG…-3' and 5'-…C^TCGAG…3', respectively. The pET24-a vector was also digested with these restriction enzymes to linearise the vector with compatible ends. Then, the ligation was performed using T4 ligase to join the compatible ends of the insert and the vector, and this ligated vector was transformed into *E. coli* DH5α cells for screening on kanamycin agar plates. The recombinant clones were confirmed by plasmid DNA sequencing using universal T7 primers. There was one nucleotide different from the reference sequence based on the *B. pseudomallei* strain K96243 genome, but the protein sequence remained unchanged (Figure 4.10).

```
ftsqh1f            1 TATACATATGCGGGAAATCCGCATCGACGGGGACACCGAGCACATCAACGCGCCGACCGT
ftsqh1r            1 TATACATATGCGGGAAATCCGCATCGACGGGGACACCGAGCACATCAACGCGCCGACCGT
BPSL3022____cel    1 -------ATGCGGGAAATCCGCATCGACGGGGACACCGAGCACATCAACGCGCCGACCGT
consensus          1 tatacatATGCGGGAAATCCGCATCGACGGGGACACCGAGCACATCAACGCGCCGACCGT

ftsqh1f           61 GCGCGCGGGCGTCGTCGGGCGGCTGAAGGGCAACTTCTTCACCGTCGATCTCGACCTCGC
ftsqh1r           61 GCGCGCGGGCGTCGTCGGGCGGCTGAAGGGCAACTTCTTCACCGTCGATCTCGACCTCGC
BPSL3022____cel   54 GCGCGCGGGCGTCGTCGGGCGGCTGAAGGGCAACTTCTTCACCGTCGATCTCGACCTCGC
consensus         61 GCGCGCGGGCGTCGTCGGGCGGCTGAAGGGCAACTTCTTCACCGTCGATCTCGACCTCGC

ftsqh1f          121 GCGCGTCGCGTTCGAGCAGATGCCGTGGGTGCGCCACGCGAGCGTGCGCCGGGTGTGGCC
ftsqh1r          121 GCGCGTCGCGTTCGAGCAGATGCCGTGGGTGCGCCACGCGAGCGTGCGCCGGGTGTGGCC
BPSL3022____cel  114 GCGCGTCGCGTTCGAGCAGATGCCGTGGGTGCGCCACGCGAGCGTGCGCCGGGTGTGGCC
consensus        121 GCGCGTCGCGTTCGAGCAGATGCCGTGGGTGCGCCACGCGAGCGTGCGCCGGGTGTGGCC

ftsqh1f          181 GAACGCGCTCGCCGTGACGCTCGAAGAGTACAAGCCGCTCGGCACGTGGGGCAACGATCA
ftsqh1r          181 GAACGCGCTCGCCGTGACGCTCGAAGAGTACAAGCCGCTCGGCACGTGGGGCAACGATCA
BPSL3022____cel  174 GAACGCGCTCGCCGTGACGCTCGAAGAGTACAAGCCGCTCGGCACGTGGGGCAACGATCA
consensus        181 GAACGCGCTCGCCGTGACGCTCGAAGAGTACAAGCCGCTCGGCACGTGGGGCAACGATCA

ftsqh1f          241 GCTCGTGAGCGTCGACGGCGAGCTCTTCACCGCGAACCAGGGCGAACTCGATGCGGAGCT
ftsqh1r          241 GCTCGTGAGCGTCGACGGCGAGCTCTTCACCGCGAACCAGGGCGAACTCGATGCGGAGCT
BPSL3022____cel  234 GCTCGTGAGCGTCGACGGCGAGCTCTTCACCGCGAACCAGGGCGAGCTCGATGCGGAGCT
consensus        241 GCTCGTGAGCGTCGACGGCGAGCTCTTCACCGCGAACCAGGGCGAaCTCGATGCGGAGCT

ftsqh1f          301 GCCGTCGTTCGACGGCCCGGAGGGCAGCGCGAAGGAGGTCGTCGCGCGCTATCGCGACTT
ftsqh1r          301 GCCGTCGTTCGACGGCCCGGAGGGCAGCGCGAAGGAGGTCGTCGCGCGCTATCGCGACTT
BPSL3022____cel  294 GCCGTCGTTCGACGGCCCGGAGGGCAGCGCGAAGGAGGTCGTCGCGCGCTATCGCGACTT
consensus        301 GCCGTCGTTCGACGGCCCGGAGGGCAGCGCGAAGGAGGTCGTCGCGCGCTATCGCGACTT

ftsqh1f          361 CGCGAAATGGTTTGCGCCGATCCACGCGACGCCTGAGGAGGTGACGCTGTCGCCGCGCTA
ftsqh1r          361 CGCGAAATGGTTTGCGCCGATCCACGCGACGCCTGAGGAGGTGACGCTGTCGCCGCGCTA
BPSL3022____cel  354 CGCGAAATGGTTTGCGCCGATCCACGCGACGCCTGAGGAGGTGACGCTGTCGCCGCGCTA
consensus        361 CGCGAAATGGTTTGCGCCGATCCACGCGACGCCTGAGGAGGTGACGCTGTCGCCGCGCTA

ftsqh1f          421 CGCGTGGACGGTGAAGCTGTCGAACGGCATGCAGGTCGAGCTCGGCCGCGAGCGCAACAG
ftsqh1r          421 CGCGTGGACGGTGAAGCTGTCGAACGGCATGCAGGTCGAGCTCGGCCGCGAGCGCAACAG
BPSL3022____cel  414 CGCGTGGACGGTGAAGCTGTCGAACGGCATGCAGGTCGAGCTCGGCCGCGAGCGCAACAG
consensus        421 CGCGTGGACGGTGAAGCTGTCGAACGGCATGCAGGTCGAGCTCGGCCGCGAGCGCAACAG

ftsqh1f          481 CGACACGCTGCCCGACCGGATCCAGCGCCTCGTCGCCGCATGGCCGTCGGTCACGCAGCG
ftsqh1r          481 CGACACGCTGCCCGACCGGATCCAGCGCCTCGTCGCCGCATGGCCGTCGGTCACGCAGCG
BPSL3022____cel  474 CGACACGCTGCCCGACCGGATCCAGCGCCTCGTCGCCGCATGGCCGTCGGTCACGCAGCG
consensus        481 CGACACGCTGCCCGACCGGATCCAGCGCCTCGTCGCCGCATGGCCGTCGGTCACGCAGCG

ftsqh1f          541 CTGGGGCGGCGACATCGAGTACGCGGATCTTCGCTATCCGAACGGATTCGCGATTCGCGC
ftsqh1r          541 CTGGGGCGGCGACATCGAGTACGCGGATCTTCGCTATCCGAACGGATTCGCGATTCGCGC
BPSL3022____cel  534 CTGGGGCGGCGACATCGAGTACGCGGATCTTCGCTATCCGAACGGATTCGCGATTCGCGC
consensus        541 CTGGGGCGGCGACATCGAGTACGCGGATCTTCGCTATCCGAACGGATTCGCGATTCGCGC

ftsqh1f          601 GGCGGGCATGCGGTTCCTGACCGATACCGACAAGGGCAAGAAGCTCGAGCACCACCACCA
ftsqh1r          601 GGCGGGCATGCGGTTCCTGACCGATACCGACAAGGGCAAGAAGCTCGAGCACCACCACCA
BPSL3022____cel  594 GGCGGGCATGCGGTTCCTGACCGATACCGACAAGGGCAAGAAGTAA--------------
consensus        601 GGCGGGCATGCGGTTCCTGACCGATACCGACAAGGGCAAGAAGctcgagcaccaccacca

ftsqh1f          661 CCACCACTGA
ftsqh1r          661 CCACCACTGA
BPSL3022____cel      ----------
consensus        661 ccaccactga
```

a)

```
ftsqh1             1 -------------------------------------MREIRIDGDTEHINAPTVRAGV
B._pseudomallei    1 MWNNVRQLNLAASALYALLLLVLAAAGCYWLIQRPAFALREIRIDGDTEHINAPTVRAGV
consensus          1 mwnnvrqlnlaasalyallllvlaaagcywliqrpafamREIRIDGDTEHINAPTVRAGV

ftsqh1            23 VGRLKGNFFTVDLDLARVAFEQMPWVRHASVRRVWPNALAVTLEEYKPLGTWGNDQLVSV
B._pseudomallei   61 VGRLKGNFFTVDLDLARVAFEQMPWVRHASVRRVWPNALAVTLEEYKPLGTWGNDQLVSV
consensus         61 VGRLKGNFFTVDLDLARVAFEQMPWVRHASVRRVWPNALAVTLEEYKPLGTWGNDQLVSV

ftsqh1            83 DGELFTANQGELDAELPSFDGPEGSAKEVVARYRDFAKWFAPIHATPEEVTLSPRYAWTV
B._pseudomallei  121 DGELFTANQGELDAELPSFDGPEGSAKEVVARYRDFAKWFAPIHATPEEVTLSPRYAWTV
consensus        121 DGELFTANQGELDAELPSFDGPEGSAKEVVARYRDFAKWFAPIHATPEEVTLSPRYAWTV

ftsqh1           143 KLSNGMQVELGRERNSDTLPDRIQRLVAAWPSVTQRWGGDIEYADLRYPNGFAIRAAGMR
B._pseudomallei  181 KLSNGMQVELGRERNSDTLPDRIQRLVAAWPSVTQRWGGDIEYADLRYPNGFAIRAAGMR
consensus        181 KLSNGMQVELGRERNSDTLPDRIQRLVAAWPSVTQRWGGDIEYADLRYPNGFAIRAAGMR

ftsqh1           203 FLTDTDKGKKLEHHHHHH
B._pseudomallei  241 FLTDTDKGKK--------
consensus        241 FLTDTDKGKKlehhhhhh
```

b)

Figure 4.10 DNA sequencing of the truncated FtsQ with a C-terminal His tag (BPSL3022_3): a) The recombinant plasmid DNA was sequenced using the universal T7 forward and reverse primers and the insert was identified by sequence alignment of the recombinant clone (read from T7 forward primer-first line, read from T7 reverse primer-second line) against the reference gene sequence (third line); b) The protein sequence of the recombinant truncated BPSL3022 (first line) was identical to the reference FtsQ protein sequence of *B. pseudomallei* strain K96243 (second line), highlighted in green. This figure was created using Boxshade (Hofmann & Baron).

BPSL3022_3 with a C-terminal His tag was expressed in *E. coli* BL21 (DE3) using 1 mM IPTG induction at 37°C for 4 hours. There was a highly expressed protein band between 21.5 and 31 kDa on the SDS PAGE gel, but only in the insoluble fraction. In order to increase the soluble form of the protein, the induction temperature was decreased to 20°C but at this temperature there was no expression of the target protein (Figure 4.11).

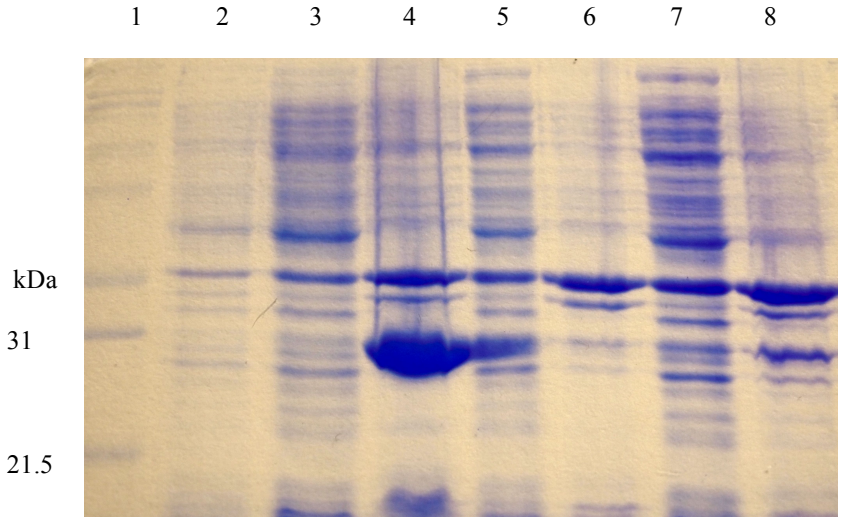


Figure 4.11 SDS PAGE of BPSL3022_3 with C-terminal His tag in pET24-a expressed in *E. coli* BL21 (DE3). lane 1 Mark12^{TM}; lane 2 pre-induction; lane 3 soluble fraction at 37°C induction for 4 hours; lane 4 insoluble fraction at 37°C induction for 4 hours; lane 5 soluble fraction at 20 °C induction 4 hours; lane 6 insoluble fraction at 20°C induction 4 hours; lane 7 soluble fraction at 20°C induction overnight; lane 8 insoluble fraction at 20°C induction overnight.

The constructs of short and long truncated FtsQ (BPSL3022_1 and BPSL3022_2) without an affinity tag were also designed and cloned into pET24-a with the restriction enzymes *Nde*I and *Bam*HI. The forward primer used to amplify these truncated genes was the same as the truncated BPSL3022 with the C-terminal His tag construct, but the reverse primers were different due to the use of different restriction enzymes and the gene fragment length. The resultant recombinant clones were confirmed by DNA sequencing and the sequence of BPSL3022_1 was identical to the reference genome in amino acid level while one amino acid

was altered in BPSL3022_2 at residue 175 (R to H). This change could be due to the different genomic strain in the database or perhaps resulted from the use of additional DMSO in the PCR. Protein expression tests were carried out using *E. coli* BL21 (DE3) with 1mM IPTG induction at 37°C for 4 hours. There was no expression of either target protein on SDS PAGE. Even if the expression tests were carried out in a different expression host *E. coli* C43 (DE3), which is able to prevent cell death associated with recombinant protein expression, but there was still no protein expression of truncated BPSL3022.

As further experiments to promote soluble expression of *B. pseudomallei* FtsQ were likely to be time consuming, this project was terminated.

4.4. BPSS0945-Subfamily M23B unassigned peptidase

The gene encoding BPSS0945 is located in chromosome 1 of the *B. pseudomallei* strain K96243 genome according to the *Burkholderia* genome. This gene consists of 948 nucleotides, encoding a 315-amino acid protein that has been annotated as a member of the subfamily M23B unassigned peptidases (Winsor *et al*, 2008).

Preliminary analysis of the BPSS0945 protein sequence was performed using ProtParam (ExPASy server) and the protein characteristics are described in Figure 4.12. BPSS0945 has a molecular weight of 32 kDa and a theoretical pI of 9.78.

```
MSKSEIDRSV AWLSGVAAAF VVAGCATTSP VTPTDTLAAA ASSASAAQPA AAPPGDAAQH
PEHAAAPAPI ATRYVVRRGD TLSAIAQANG CTVRELQAWN RMGRRTRIGI GQVLRIAPPG
AENAVASEAA DGVGATRAAG NAAGAARAAD PASGAGGAPV ADAPQPASAP ESAADRAADH
RVVQETKRHA QSIALAWPAK GAVVETFQPG RNRGIRIVGR AGDPVRAAAS GRVMYAGTGL
NGYGTLILVQ HNADFLTAYA HNRKVLVKTG DVVQQGEQIA EMGTGDSTRA GMLFEVRRDG
KPVNPMPYLA SRQQG
```

Number of amino acids: 315 **Molecular weight:** 32027.9 **Theoretical pI:** 9.78

Amino acid composition:

Ala (A)	69	21.9%	Ile (I)	11	3.5%	Tyr (Y)	5	1.6%
Arg (R)	26	8.3%	Leu (L)	13	4.1%	Val (V)	26	8.3%
Asn (N)	9	2.9%	Lys (K)	6	1.9%	Pyl (O)	0	0.0%
Asp (D)	14	4.4%	Met (M)	6	1.9%	Sec (U)	0	0.0%
Cys (C)	2	0.6%	Phe (F)	4	1.3%			
Gln (Q)	15	4.8%	Pro (P)	21	6.7%	(B)	0	0.0%
Glu (E)	11	3.5%	Ser (S)	17	5.4%	(Z)	0	0.0%
Gly (G)	33	10.5%	Thr (T)	18	5.7%	(X)	0	0.0%
His (H)	6	1.9%	Trp (W)	3	1.0%			

Total number of negatively charged residues (Asp + Glu): 25
Total number of positively charged residues (Arg + Lys): 32

Atomic composition:

Carbon	C	1375
Hydrogen	H	2224
Nitrogen	N	438
Oxygen	O	430
Sulfur	S	8

Formula: $C_{1375}H_{2224}N_{438}O_{430}S_8$ **Total number of atoms:** 4475

Extinction coefficients:
Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.
Ext. coefficient 24075
Abs 0.1% (=1 g/l) 0.752, assuming ALL Cys residues appear as half cystines

Figure 4.12 Analysis of BPSS0945 protein sequence: The characterization of BPSS0945 revealed its molecular weight, theoretical pI, amino acid composition and extinction coefficients by ProtParam.

The hydrophobicity of BPSS0945 was also characterised by hydropathy plot (http://www.vivo.colostate.edu/molkit/hydropathy/index.html). Kyte-Doolittle scale plot of BPSS0945 showed the protein is rich in hydrophobic residues at the N-terminal region (Figure 4.13). It is unlikely to be a transmembrane protein but this hydrophobic region could possibly be a signal peptide.

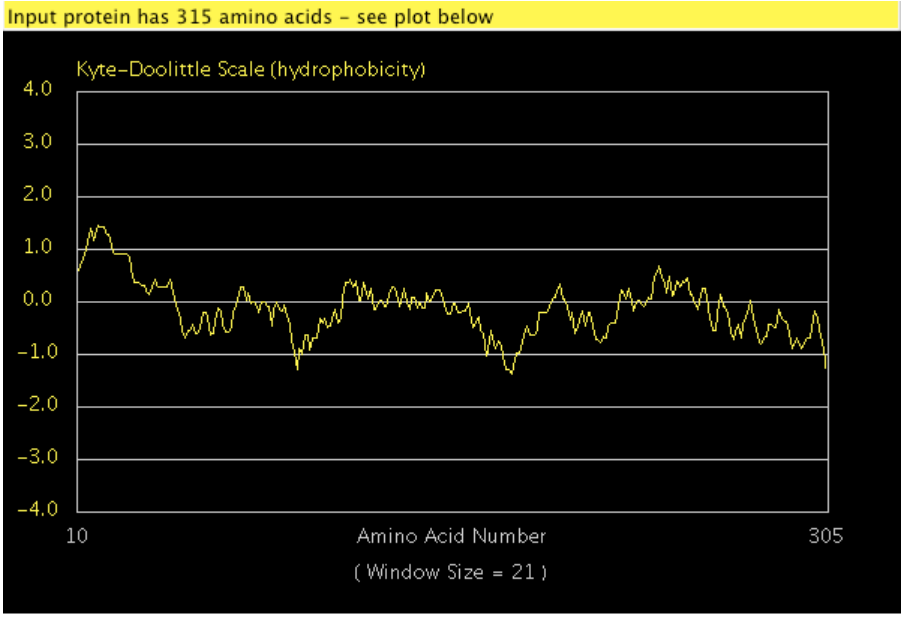
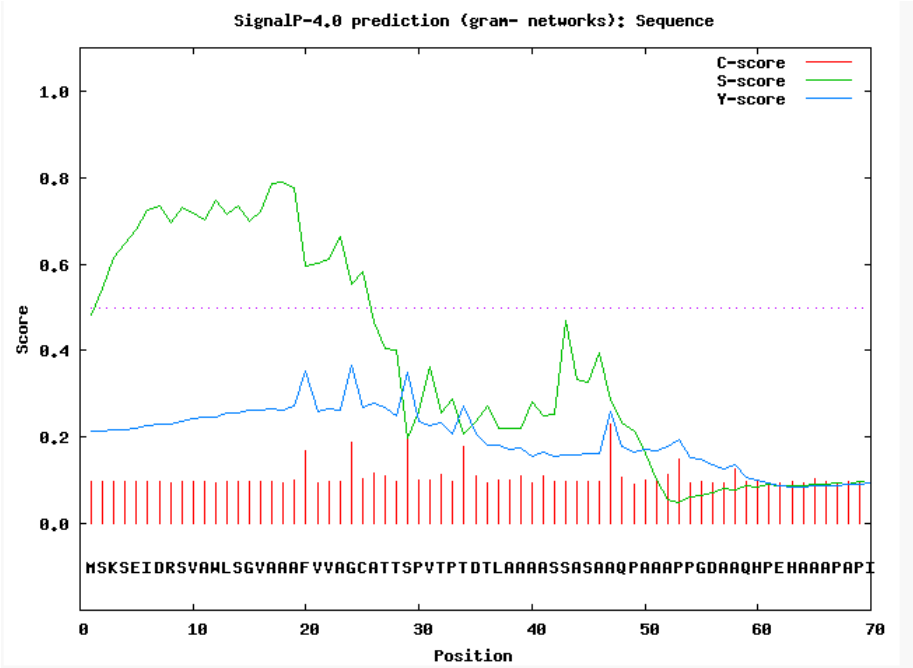


Figure 4.13 A hydropathy plot of BPSS0945: The N-terminal region of the protein contains hydrophobic residues. This figure was produced using the hydropathy plot server (http://www.vivo.colostate.edu/molkit/hydropathy/index.html).

To predict whether a signal peptide sequence was present in BPSS0945, the protein sequence was analysed using the SignalP 4.0 server (Petersen *et al*, 2011). According to the scores, it was likely that residues 1-23 are probably a signal peptide, but this annotation is not certain.



# Measure	Position	Value	Cutoff	signal peptide?
max. C	47	0.229		
max. Y	24	0.365		
max. S	18	0.791		
mean S	1-23	0.684		
D	1-23	0.515	0.570	NO

Figure 4.14 Signal peptide prediction of BPSS0945: There is no predicted signal peptide in BPSS0945 using the score cut-off at 0.57. This figure was produced using the SignalP 4.0 server (Petersen *et al*, 2011).

4.4.1. Protein structure prediction

Protein structure prediction for fold recognition and secondary structure for BPSS0945 was performed using the PSIPRED Protein Structure Prediction Server (McGuffin *et al*, 2000). BPSS0945 was matched to three proteins in the PDB, with more than 99% confidence level. These proteins were a lysostaphin peptidase from *Vibrio cholerae*, a peptidoglycan hydrolase from *Staphylococcus aureus* and an M23 family peptidase from *Pseudomonas aeruginosa* (Table 4.4).

Table 4.4 A list of similar protein structures to the predicted BPSS0945 structure

PDB ID	Protein	Organism	% Sequence identity to BPSS0945	Crystallisation condition
2gu1	Lysostaphin peptidase	*V. cholerae*	36	0.2M $NH_2(SO_4)$, 20% PEG 4000, 0.1 Tris pH 8.5, 25°C
1qwy	Peptidoglycan hydrolase, LytM	*S. aureus*	32	0.17M $NH_2(SO_4)$, 25.5%(w/v)PEG 8K, 15% (v/v) glycerol, 21°C
2hsi	M23 peptidase	*P. aeruginosa*	30	0.1M HEPES, 0.2M CH_3COONH_4, 25% PEG 3350, 10% glycerol, pH 6.5, 17°C

The protein sequences of BPSS0945 and the identified homologue proteins were aligned and the key residues were identified. As the signature-conserved sequences commonly found in these lysostaphin-type like enzymes, which are metallopeptidases, have been described, the motifs involved in metal ligand binding were considered in this case. The motifs HX(3,6)D and HXH have been reported in which the ligand binding occurs. The residues that chelate the metal, which are H and D in the HX(3,6)D motif and the first H in the HXH motif are conserved in space in all LytM crystal forms, and found in the Zn^{2+} binding site in peptidoglycan amidases (Firczuk *et al*, 2005). The sequence alignment shows that, despite high sequence similarity, BPSS0945 lacks these two conserved motifs relating to Zn^{2+} binding (Figure 4.15). It is possible that BPSS0945 either does not bind metal or uses a different constellation of residues than those seen in LytM-like proteins.

```
1QWY_A_PDBID_CH     1 --------------AETTNTQQAHT------QMSTQSQDVSYGTYYTIDSNGDYHHTPD-
2GU1_A_PDBID_CH     1 MSLQPKRIHYMVKVGDTLSGIFAQLGVPYSILQKILSVDLDHLQLDMIQPGEELELMMDD
BPSS0945_gi_522     1 --MSKSEIDRSVAWLSGVAAAFVVAG--CATTSPVTPTDTLAAAASSASAAQPAAAPPGD
M23_Pseudomonas     1 -------MPRTLAFVSTLLLAAFCA-------LPAQADSFIMRLLNKPVPGGVAVVDLGE
consensus           1 --l----i-rsvaw-stl-gafa-ag------mpiqs-dvl-g-l-ti-pggda-l-p-d

1QWY_A_PDBID_CH    40 -GNWNQAMFDNKEYSYTFVDAQGHTHYFYNCYPKN-------ANANGSGQTYVNPATAGD
2GU1_A_PDBID_CH    61 MGQLSRLIYHMSIVEKAIYTRENDGSFSYDFQEISGEWREILFSGEINGSFSVSARRVGL
BPSS0945_gi_522    57 AAQHPEHAAAPAPIATRYVVRRGDT--LSAIAQANGCTVRELQAWNRMGRRTRIGIGQVL
M23_Pseudomonas    47 EGPPPRAFYQGK---PVLVVRE-------------------------EGRRWIAVVGIPL
consensus          61 -gq-pramy--k-v----vvregdt-y-y-----ng-----l--an--Grr-v-aigvgl

1QWY_A_PDBID_CH    92 NNDYTASQSQQHINQYGYQSNVGPDASYYSHSNNNQAYNSHDGNGKVNYP----------
2GU1_A_PDBID_CH   121 TSSQVANITQVMKDKIDFSRSLRAGDRFDILVKQQYLGEHNTGNSEIKAISFKLAKGDVS
BPSS0945_gi_522   115 RIAPPGAENAVASEAADGVGATRAAGNAAGAARAADPASGAGGAPVADAP----------
M23_Pseudomonas    79 STKPGPQKLEVRAATGNHEERFSVGSKHYREQRITLKNKRQVNPLPEDLKRIER------
consensus         121 ts-p-an-sqv--d-ady---vragaryy---rnn--g--n-gn--vdap----------

1QWY_A_PDBID_CH   142 -----NGTSNQNGGSASKATASGHAKDASWLTSRKQLQPYGQYHGGG---AHYGVDYAMP
2GU1_A_PDBID_CH   181 AFLAEDGRFYDRAGNSLERAFNRYPVDKAYRQITSGFNPKRKHPVTGRVVPHNGTDFATP
BPSS0945_gi_522   165 ----QPASAPESAADRAADHRVVQETKRHAQSIALAWPAKGAVVETFQPGRNRGIRIVGR
M23_Pseudomonas   133 ----ELAEQTAAYRRFSPGLPSNLMLDKPVDGPLSSPFGLRRFFNGEERNPHSGLDFAVP
consensus         181 ----e-gt--d-ag--s-a--s---vdk-w-ti-sgfqpk-ky---g---ph-Gvdfamp

1QWY_A_PDBID_CH   194 ENSPVYSLTDGTVVQAGWSNYGGGNQVTIKEANSNNYQWYMHNNRLTVSAGDKVKAGDQI
2GU1_A_PDBID_CH   241 IGAPVYSTGDGKVIVVRKHPYAG--NYLVIEHNSVYKTRYLHLDKILVKKGQLVKRGQKI
BPSS0945_gi_522   221 AGDPVRAAASGRVMYAGTG-LNGYGTLILVQHNADFLTAYAHNRKVLVKTGDVVQQGEQI
M23_Pseudomonas   189 AGTPIKAPAAGKVILIGDYFFNG--KTVFVDHGQGFISMFCHLSKIDVKLGQQVPRGGVL
consensus         241 agsPvy--adGkVivag---ynG--qvlivehns-flt-ymH--kilVk-G-lVkrGdqi

1QWY_A_PDBID_CH   254 AYSGSTGNSTAPHVHFQRMSGG-----------IGNQYAVDPTSYLQSR-----------
2GU1_A_PDBID_CH   299 ALAGATGRLTGPHLHFEVLVRNRPVDAMKADLPIAKSLSSNQKTSFLARVSEFDHEGHHH
BPSS0945_gi_522   280 AEMG-TGDSTRAGMLFEVRRDG---------------KPVNPMPYLASRQQG--------
M23_Pseudomonas   247 GKVGATGRATGPHMHWNVSLND---------------ARVDPAIFIGAFQP---------
consensus         301 a-mGaTGrsTgphmhfevmv-g-----------ig----v-p-syla-rq----------

1QWY_A_PDBID_CH       ---
2GU1_A_PDBID_CH   359 HHH
BPSS0945_gi_522       ---
M23_Pseudomonas       ---
consensus         361 ---
```

Figure 4.15 Protein sequence alignment of BPSS0945 (third line) and proteins with known structure; peptidoglycan hydrolase (LytM) from *S. aureus* (first line), Lysostaphin peptidase from *V. cholerae* (second line) and M23 peptidase from *P. aeruginosa* (fourth line): Identical residues are displayed in green boxes. Metal ligand binding motifs are shown with a red border. This figure was created using Boxshade (Hofmann & Baron).

4.4.2. Gene amplification and cloning

To clone the BPSS0945 protein, oligonucleotide primers were designed according to the open reading frame of the BPSS0945 coding gene, subject to being cloned into a pETBlue-1 vector. The forward primer 5'-ATGAGCAAGAGCGAGAT CGA-3' and the reverse primer 5'- CTAGCCCTGCTGCCGGC-3' were used together with the genomic DNA of *B. pseudomallei* strain D286 to amplify the gene fragment by PCR. The PCR product was obtained with the addition of 5% DMSO at an annealing temperature of 56°C. Analysis on 1% agarose gel electrophoresis showed that the PCR product was about 1,000 base pairs in size, and was likely to be BPSS0945 which consists of 948 base pairs (Figure 4.16).

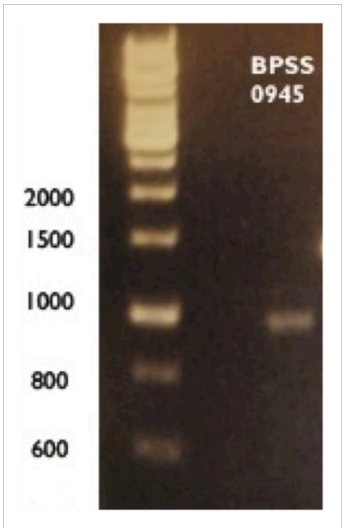


Figure 4.16 Amplification of BPSS0945 coding gene: A 1% agarose gel showing that PCR product has approximately 1000 base pairs, corresponding to BPSS0945 in size (948 base pairs).

The PCR product of BPSS0945 was cloned into the pETBlue-1 vector using the TA cloning strategy. After transformation of *E. coli* NovaBlue, which facilitates blue/white screening, the white colonies were selected and verified for correctly oriented insert using colony PCR, prior to DNA sequencing. The DNA sequencing results of BPSS0945 recombinant clones showed there were three nucleotides different from the *B. pseudomallei* K96243 genome sequence. As a result of these nucleotide differences, 2 amino acids were different in the protein sequence, residue 133 (V to A) and residue 177 (A to V) (Figure 4.17). In the sequences of orthologous proteins from different strains of other completely sequenced genomes of *B. pseudomallei*, the equivalent residues of the altered residue 133 are all valine, whereas those of the residue 177 are either alanine or valine. These observations could possibly be due to a result of a mutation in the PCR process or a real strain variation. A further clarification of the sequence differences is required.

```
0945r        1 --------------------------NNNNNNTTTGTTNAACTTTANGNNGGAGATATAGATTATGAGCAAGAGCGAGAT
BPSS0945     1 ------------------------------------------------------------------ATGAGCAAGAGCGAGAT
0945f        1 NNNNNNNNNNNNNNTNNTTGNCTAGCAGATAATTTTGTTTAACTTTAAGAAGGAGATATAGATTATGAGCAAGAGCGAGAT
consensus    1 ------------------------------tttgtt-aacttta-g--ggagatatagattATGAGCAAGAGCGAGAT

0945r       55 CGACCGCAGTGTCGCATGGTTGTCCGGCGTGGCGGCCGCGTTCGTCGTCGCGGGCTGCGCAACCACGTCGCCCGTCACGC
BPSS0945    18 CGACCGCAGTGTCGCATGGTTGTCCGGCGTGGCGGCCGCGTTCGTCGTCGCGGGCTGCGCAACCACGTCGCCCGTCACGC
0945f       81 CGACCGCAGTGTCGCATGGTTGTCCGGCGTGGCGGCCGCGTTCGTCGTCGCGGGCTGCGCAACCACGTCGCCCGTCACGC
consensus   81 CGACCGCAGTGTCGCATGGTTGTCCGGCGTGGCGGCCGCGTTCGTCGTCGCGGGCTGCGCAACCACGTCGCCCGTCACGC

0945r      135 CGACTGATACGCTCGCGGCCGCCGCGTCTTCGGCGAGCGCCGCGCAGCCCGCGGCCGCGCCGCCGGGCGACGCGGCGCAG
BPSS0945    98 CGACTGATACGCTCGCGGCCGCCGCGTCTTCGGCGAGCGCCGCGCAGCCCGCGGCCGCGCCGCCGGGCGACGCGGCGCAG
0945f      161 CGACTGATACGCTCGCGGCCGCCGCGTCTTCGGCGAGCGCCGCGCAGCCCGCGGCCGCGCCGCCGGGCGACGCGGCGCAG
consensus  161 CGACTGATACGCTCGCGGCCGCCGCGTCTTCGGCGAGCGCCGCGCAGCCCGCGGCCGCGCCGCCGGGCGACGCGGCGCAG

0945r      215 CATCCGGAGCACGCCGCCGCGCCGGCGCCGATCGCGACACGCTACGTCGTGAGGCGCGGCGATACGCTGTCGGCGATCGC
BPSS0945   178 CATCCGGAGCACGCCGCCGCGCCGGCGCCGATCGCGACACGCTACGTCGTGAGGCGCGGCGATACGCTGTCGGCGATCGC
0945f      241 CATCCGGAGCACGCCGCCGCGCCGGCGCCGATCGCGACACGCTACGTCGTGAGGCGCGGCGATACGCTGTCGGCGATCGC
consensus  241 CATCCGGAGCACGCCGCCGCGCCGGCGCCGATCGCGACACGCTACGTCGTGAGGCGCGGCGATACGCTGTCGGCGATCGC

0945r      295 GCAAGCGAACGGCTGCACCGTGCGCGAGCTGCAGGCCTGGAACCGGATGGGCCGGCGCACCCGGATCGGCATCGGCCAGG
BPSS0945   258 GCAAGCGAACGGCTGCACCGTGCGCGAGCTGCAGGCTTGGAACCGGATGGGCCGGCGCACCCGGATCGGCATCGGCCAGG
0945f      321 GCAAGCGAACGGCTGCACCGTGCGCGAGCTGCAGGCCTGGAACCGGATGGGCCGGCGCACCCGGATCGGCATCGGCCAGG
consensus  321 GCAAGCGAACGGCTGCACCGTGCGCGAGCTGCAGGCcTGGAACCGGATGGGCCGGCGCACCCGGATCGGCATCGGCCAGG

0945r      375 TATTGCGGATCGCGCCGCCGGGTGCGGAAAACGCGGTCGCCTCCGAGGCGGCGGATGGCGCGGGCGCGACCCGCGCGGCC
BPSS0945   338 TATTGCGGATCGCGCCGCCGGGTGCGGAAAACGCGGTCGCCTCCGAGGCGGCGGATGGCGTGGGCGCGACCCGCGCGGCC
0945f      401 TATTGCGGATCGCGCCGCCGGGTGCGGAAAACGCGGTCGCCTCCGAGGCGGCGGATGGCGCGGGCGCGACCCGCGCGGCC
consensus  401 TATTGCGGATCGCGCCGCCGGGTGCGGAAAACGCGGTCGCCTCCGAGGCGGCGGATGGCGcGGGCGCGACCCGCGCGGCC

0945r      455 GGCAATGCAGCCGGCGCCGCGCGCGCGGCCGATCCGGCGAGCGGCGCGGGCGGGGCGCCTGTCGCCGATGCGCCGCAGCC
BPSS0945   418 GGCAATGCAGCCGGCGCCGCGCGCGCGGCCGATCCGGCGAGCGGCGCGGGCGGGGCGCCTGTCGCCGATGCGCCGCAGCC
0945f      481 GGCAATGCAGCCGGCGCCGCGCGCGCGGCCGATCCGGCGAGCGGCGCGGGCGGGGCGCCTGTCGCCGATGCGCCGCAGCC
consensus  481 GGCAATGCAGCCGGCGCCGCGCGCGCGGCCGATCCGGCGAGCGGCGCGGGCGGGGCGCCTGTCGCCGATGCGCCGCAGCC

0945r      535 TGCATCGGCGCCCGAAAGCGCCGCGGACCGGGTGGCGGACCATCGGGTCGTCCAGGAAACGAAGCGGCACGCGCAATCGA
BPSS0945   498 TGCATCGGCGCCCGAAAGCGCCGCGGACCGGGCGGCGGACCATCGGGTCGTCCAGGAAACGAAGCGGCACGCGCAATCGA
0945f      561 TGCATCGGCGCCCGAAAGCGCCGCGGACCGGGTGGCGGACCATCGGGTCGTCCAGGAAACGAAGCGGCACGCGCAATCGA
consensus  561 TGCATCGGCGCCCGAAAGCGCCGCGGACCGGGtGGCGGACCATCGGGTCGTCCAGGAAACGAAGCGGCACGCGCAATCGA

0945r      615 TCGCGCTCGCGTGGCCGGCGAAGGGCGCGGTCGTCGAAACGTTCCAGCCCGGCCGCAATCGCGGCATCCGGATCGTCGGG
BPSS0945   578 TCGCGCTCGCGTGGCCGGCGAAGGGCGCGGTCGTCGAAACGTTCCAGCCCGGCCGCAATCGCGGCATCCGGATCGTCGGG
0945f      641 TCGCGCTCGCGTGGCCGGCGAAGGGCGCGGTCGTCGAAACGTTCCAGCCCGGCCGCAATCGCGGCATCCGGATCGTCGGG
consensus  641 TCGCGCTCGCGTGGCCGGCGAAGGGCGCGGTCGTCGAAACGTTCCAGCCCGGCCGCAATCGCGGCATCCGGATCGTCGGG

0945r      695 CGCGCGGGCGATCCGGTGCGTGCCGCGGCGTCCGGCCGGGTGATGTACGCGGGCACGGG-CCTGAACGGGTACGGCACGC
BPSS0945   658 CGCGCGGGCGATCCGGTGCGTGCCGCGGCGTCCGGCCGGGTGATGTACGCGGGCACGGG-CCTGAACGGGTACGGCACGC
0945f      721 CGCGCGGGCGATCCGGTGCGTGCCGCGGCGTCCGGCCGGGTGATGTACGCGGGCACGGGCCTGAACGGGTACGGCACGCT
consensus  721 CGCGCGGGCGATCCGGTGCGTGCCGCGGCGTCCGGCCGGGTGATGTACGCGGGCACGGG-CctgAacGGgtacGgcacgc

0945r      774 TGATCCTGGTCCAGCACAACGCGGATTTTCTGACCGCTTACGCGCACAACCG--CAAGGTGCTCGTGAAGACGGGCGACG
BPSS0945   737 TGATCCTGGTCCAGCACAACGCGGATTTTCTGACCGCTTACGCGCACAACCG--CAAGGTGCTCGTGAAGACGGGCGACG
0945f      801 GATCCTGGTCCAGCACAACGCGGATTTTCTGACCGCTTACGCGCACAACCGCAAGGTGCTCGTGAAGACGGGCGACGTCG
consensus  801 tgatCctGgtCcagcacAacgcGgaTTTtctgaCcgcTtacgcgcacAaCcg--caaGgTgcTcgtGAaGacgGgCgaCG

0945r      852 TCGTTCAGC-AGGGCGAACAGATCGCCGAGATGGGCACCGGCGACAGCACGCGCCGGCATGCTGTTCGAGGTGCGGCG
BPSS0945   815 TCGTTCAGC-AGGGCGAACAGATCGCCGAGATGGGCACCGGCGACAGCACGCGCGCCGGCATGCTGTTCGAGGTGCGGCG
0945f      881 TTCAAGCAGGGCGAACAGATCNCCCAGATGGGCACCGGC-------------------------------------
consensus  881 Tcgttcagc-agGgcgAacagatCgccgaGatgggCacCggcgacagcacgcgcgccggcatgctgttcgaggtgcggcg

0945r      931 CGACGGCAAGCCCGTCAATCCGATGCCGTATCTGGCGAGCCGGCAGCAGGGCTAGAATCGAATTCCTGCCCGGGCGTGTA
BPSS0945   894 CGACGGCAAGCCCGTCAATCCGATGCCGTATCTGGCGAGCCGGCAGCAGGGCTAG-------------------------
0945f      ------------------------------------------------------------------------------------
consensus  961 cgacggcaagcccgtcaatccgatgccgtatctggcgagccggcagcagggctag-------------------------
```

a)

```
BPSS0945   MSKSEIDRSVAWLSGVAAAFVVAGCATTSPVTPTDTLAAAASSASAAQPAAAPPGDAAQH
25_0945    MSKSEIDRSVAWLSGVAAAFVVAGCATTSPVTPTDTLAAAASSASAAQPAAAPPGDAAQH
consensus  MSKSEIDRSVAWLSGVAAAFVVAGCATTSPVTPTDTLAAAASSASAAQPAAAPPGDAAQH

BPSS0945   PEHAAAPAPIATRYVVRRGDTLSAIAQANGCTVRELQAWNRMGRRTRIGIGQVLRIAPPG
25_0945    PEHAAAPAPIATRYVVRRGDTLSAIAQANGCTVRELQAWNRMGRRTRIGIGQVLRIAPPG
consensus  PEHAAAPAPIATRYVVRRGDTLSAIAQANGCTVRELQAWNRMGRRTRIGIGQVLRIAPPG

BPSS0945   AENAVASEAADGVGATRAAGNAAGAARAADPASGAGGAPVADAPQPASAPESAADRAADH
25_0945    AENAVASEAADGAGATRAAGNAAGAARAADPASGAGGAPVADAPQPASAPESAADRVADH
consensus  AENAVASEAADG-GATRAAGNAAGAARAADPASGAGGAPVADAPQPASAPESAADR-ADH

BPSS0945   RVVQETKRHAQSIALAWPAKGAVVETFQPGRNRGIRIVGRAGDPVRAAASGRVMYAGTGL
25_0945    RVVQETKRHAQSIALAWPAKGAVVETFQPGRNRGIRIVGRAGDPVRAAASGRVMYAGTGL
consensus  RVVQETKRHAQSIALAWPAKGAVVETFQPGRNRGIRIVGRAGDPVRAAASGRVMYAGTGL

BPSS0945   NGYGTLILVQHNADFLTAYAHNRKVLVKTGDVVQQGEQIAEMGTGDSTRAGMLFEVRRDG
25_0945    NGYGTLILVQHNADFLTAYAHNRKVLVKTGDVVQQGEQIAEMGTGDSTRAGMLFEVRRDG
consensus  NGYGTLILVQHNADFLTAYAHNRKVLVKTGDVVQQGEQIAEMGTGDSTRAGMLFEVRRDG

BPSS0945   KPVNPMPYLASRQQG
25_0945    KPVNPMPYLASRQQG
consensus  KPVNPMPYLASRQQG
```

b)

Figure 4.17 DNA sequencing of the BPSS0945 recombinant clone: The recombinant plasmid DNA was sequenced and the insert was identified; a) Nucleotide sequence alignment of BPSS0945 (second line) and positive clone sequenced using T7 forward primer (third line) and T7 reverse primer (first line) shows three nucleotides in the recombinant gene which had been amplified from *B. pseudomallei* strain D286 were different from the reference sequence of *B. pseudomallei* strain K96243; b) Protein sequence alignment of the reference sequence of BPSS0945 (first line) and the identified insert sequence from positive clone (second line) shows two amino acid residues altered (colored in red) from the reference sequence of *B. pseudomallei* strain K96243.

4.4.3. Protein expression

To produce recombinant protein of BPSS0945, *E. coli* Tuner^{TM} (DE3) cells were used to express the gene encoding BPSS0945, which had been inserted in a pETBlue-1 vector. The recombinant protein was over-produced by adding 1mM IPTG. After induction at 37°C for 4 hours, SDS PAGE showed that there were obvious bands around 36.5 kDa in both the soluble and insoluble fractions, indicative of the target protein, which has the approximate molecular weight of 32 kDa (Figure 4.18). Reducing the expression temperature to 25°C did not increase the soluble protein expression and at 20°C there was no expression of the target protein.

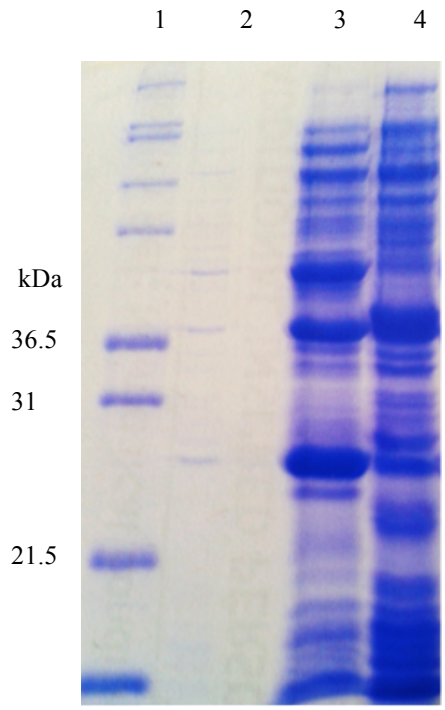


Figure 4.18 Overexpression of BPSS0945: SDS PAGE of BPSS0945 expressed in *E.coli* Tuner^{TM} (DE3) pLacI on 12% resolving gel; lane 1 a protein marker Mark12^{TM}; lane 2 pre-induction; lane 3 soluble fraction at 37°C induction for 4 hours; lane 4 insoluble fraction at 37°C induction for 4 hours.

4.4.4. Protein purification

About 9 g of cell paste were resuspended in 50 mM HEPES pH 7.5 and the cells were disrupted by ultra-sonication on ice for 2-3 cycles of 20 seconds. Crude protein was extracted and separated by centrifugation at 24500 rpm for 10 minutes. Approximately 56 ml of crude extract was loaded onto a 5 ml SP-HP cation exchange column (GE Healthcare), which was pre-equilibrated with 50 mM HEPES pH 7.5. The protein was eluted with a gradient of 0-1 M NaCl in 50 mM HEPES pH 7.5. 3 ml fractions were collected and monitored by UV absorbance. The protein concentration of fractions corresponding to the UV peak was determined by the Bradford assay and the fraction profile was analysed by SDS PAGE. The fractions containing BPSS0945 (32 kDa) were combined. Ammonium sulphate precipitation was subsequently performed by adding 4 M ammonium sulphate to make a final concentration of 1.5 M ammonium sulphate. The protein precipitate was recovered by centrifugation at 24500 rpm for 5-10 minutes and resuspended with 50 mM HEPES pH 7.5 prior to the next purification step using gel filtration. The protein sample was loaded onto Superdex™ 200 (GE Healthcare). The protein was eluted using 50 mM Tris-HCl pH 8 plus 0.5 M NaCl. 2 ml fractions were collected and monitored by UV absorbance. The protein concentration of fractions corresponding to the UV peak was determined by the Bradford assay and the fraction profile was analysed by SDS PAGE. The fractions containing the purified BPSS0945 protein were pooled together (Figure 4.19) and the protein was concentrated to approximately 10 mg/ml in 10 mM HEPES pH 7.5 with 50 mM NaCl using a 10 kDa MWCO Vivaspin™ centrifugal concentrator (Vivascience). The average yield of protein purification was about 3 mg purified protein obtained from three litres of recombinant *E. coli* culture.

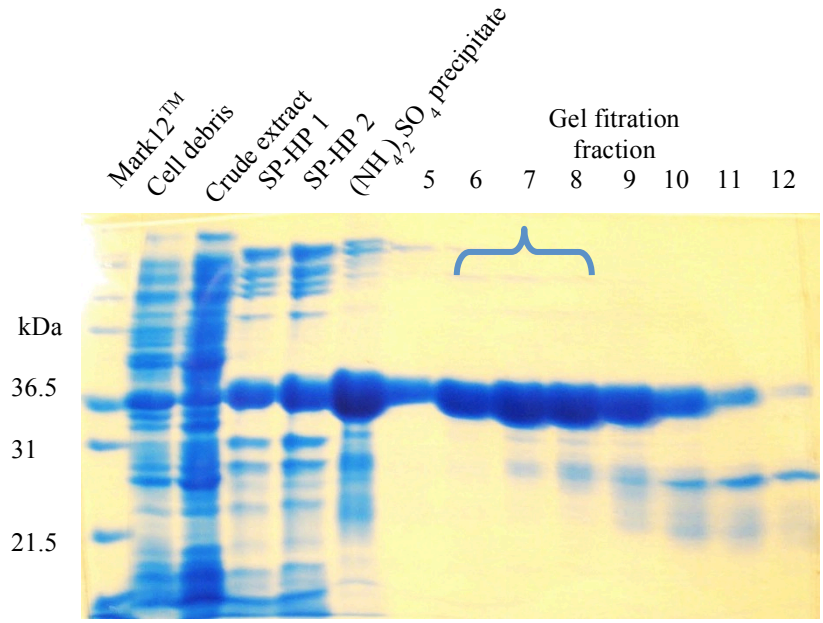


Figure 4.19 Purification of BPSS0945: SDS PAGE gel shows the recovery of the target protein BPSS0945 in each step of protein purification. Fraction 6, 7 and 8 in the gel filtration step were collected and the protein was concentrated prior to crystallisation trials.

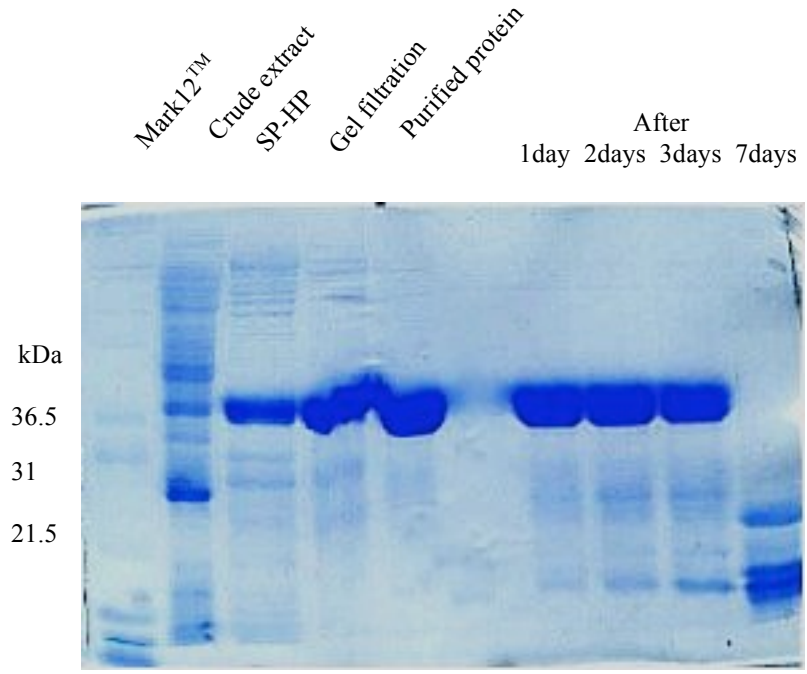


Figure 4.20 Stability of BPSS0945: SDS PAGE gel shows the BPSS0945 protein degradation after purification and stored at 4 °C for up to 7 days.

4.4.5. Crystallisation trials

The purified protein was concentrated to approximately 10 mg/ml and then put into crystallisation trials. Initial crystallisation screening was performed using commercially available NeXtal crystallisation screens (QIAGEN) by the sitting drop vapour diffusion technique. 9.5 mg/ml purified protein in 50 mM HEPES pH 7 were screened using a Hydra II Plus 1 robot against JCSG, PACT and Classic screens. No crystals were observed and most drops were clear, indicating a high solubility. Therefore, the protein concentration was increased to 14.5 mg/ml protein and used against the same screens. Yet again, no crystals were obtained.

4.4.6. Unstable BPSS0945

Due to the crystallisation failure, the stability of BPSS0945 was analysed after purification. SDS PAGE showed the degradation of protein occurs over time, possibly due to its peptidase activity (Figure 4.20). The full-length protein completely disappeared within 7 days. This non-homogenous protein solution could easily cause failure in the crystallisation step.

The same autolytic phenomenon also occurred in the production of the related peptidase LytM, a unique autolysin from *Staphylococcus aureus* (Ramadurai & Jayaswal, 1997). The autolytic activity bands of degraded products of LytM were 19 and 22 kDa. This protein was further characterised and it may possibly be a glycylglycine endopeptidase, based on the biochemical analysis of LytM-cleaved peptidoglycan fragments (Ramadurai *et al*, 1999). BPSS0945 shares 32% sequence identity with LytM and it is possible that a similar autolysis to LytM is seen for BPSS0945.

As the sequence alignments showed that BPSS0945 is possibly an M23B metal-dependent peptidase family member, attempts were made to inactivate its peptidase function by the addition of 4 mM EDTA, to act as a chelating agent for the metal ion.

Purified protein at 21 mg/ml with EDTA treatment was again put into screening crystallisation trials but there were still no crystals. The protein treated with EDTA was analysed on SDS gel showing that the degradation of BPSS0945 still occurred (Figure 4.21). However, as the Zn^{2+} binding motif present in the LytM family cannot be found in the sequence of BPSS0945, it is possible that the peptidase function of this protein does not require metal, and hence EDTA treatment would be ineffective.

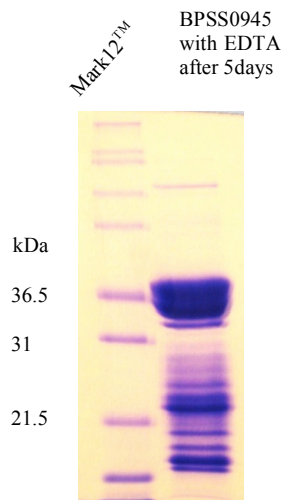


Figure 4.21 BPSS0945 treated with EDTA: In order to inactivate the activity of BPSS0945, EDTA was added to the protein after purified. It seems that BPSS0945 still has activity to cleave itself, resulting in non-homogenous solution, which may cause crystallisation failure.

4.5. BPSS0238 and BPSS1416

4.5.1. BPSS0238 Penicillin-binding protein

The gene encoding BPSS0238 is located in chromosome 2 of *B. pseudomallei* strain K96243. This gene consists of 2163 nucleotides, encoding a 720-amino acid protein. According to the *Burkholderia* genome database, it has been annotated as a putative penicillin-binding protein (Winsor *et al*, 2008). The protein sequence of BPSS0238 was analysed by ProtParam (ExPasy server). This protein has a molecular weight of approximately 77 kDa with a theoretical pI of 9.4 (Figure 4.22).

```
MPTKDFDMTD HTPHSSPPAS PTPPQGPAPK RRRYFRTLAG LVLGAAIAGG
AVGTWTLHRI WTALPSVEPL AVYRPALPLR IFSREGELLA EYGVERREFV
PLDRIPPLVR HALLAAEDAQ FYTHGVVDIG GLARATVANV VTGQPGQGGS
TITMQVARNF FLTREKVLSR KLAEILMSYK LERAYSKDKL LELYMNEIYL
GERAYGFAAA ASVYFGKPLD ALTPGEAAVL AGLPKAPSAF NPVVNPARAT
MRRNYVLGRM RALGYLDEAA YRQAADAPIA LATTPPPGIL AAPYVAERAR
RMMVERFHDD AYTLGLDVTT TIAMREQRAA DAALARGLAQ VGRRDAKDAL
EGALVSIDAA TGDILALVGG ADFSHNVFDH ALQAYRQPGS SFKPFVYSAA
LEKGMFPGVL IDDTQRTLSR DETGANPWRP RNFGNHYEGF IPVRRGLMRS
KNLVAVSLMQ VARPDFVQQH AARFGFEPLR NPASLPLALG AGSATPLEVA
SAYGVFANGG VRMEPRLIES VRQRHGGALF DAQPAPGVRV ISARNAFVMD
SMLRDVVAHG TARRAVALKR ADAAGKTGTS NGSKDVWFAG YSSGVVSVVW
MGYDAPRSLG RATGSSAALP VWVDYMKTAV DGRGAIERTP PADVALVDGD
FVYAEYANGD KCAPSLPPFV RSPFACGAAR GAPADGGAPA SAAGAAGAPT
PAAVDAAERA RVLELFRTED
```

Number of amino acids: 720 **Molecular weight:** 76876.7 **Theoretical pI:** 9.40

Amino acid composition:

Ala (A)	118	16.4%	Ile (I)	18	2.5%	Tyr (Y)	24	3.3%
Arg (R)	63	8.8%	Leu (L)	63	8.8%	Val (V)	57	7.9%
Asn (N)	16	2.2%	Lys (K)	18	2.5%	Pyl (O)	0	0.0%
Asp (D)	39	5.4%	Met (M)	18	2.5%	Sec (U)	0	0.0%
Cys (C)	2	0.3%	Phe (F)	29	4.0%			
Gln (Q)	16	2.2%	Pro (P)	54	7.5%	(B)	0	0.0%
Glu (E)	31	4.3%	Ser (S)	36	5.0%	(Z)	0	0.0%
Gly (G)	65	9.0%	Thr (T)	35	4.9%	(X)	0	0.0%
His (H)	12	1.7%	Trp (W)	6	0.8%			

Total number of negatively charged residues (Asp + Glu): 70
Total number of positively charged residues (Arg + Lys): 81

Atomic composition:

Carbon	C	3425
Hydrogen	H	5395
Nitrogen	N	989
Oxygen	O	988
Sulfur	S	20

Formula: $C_{3425}H_{5395}N_{989}O_{988}S_{20}$
Total number of atoms: 10817

Extinction coefficients:

Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

Ext. coefficient 68885
Abs 0.1% (=1 g/l) 0.896, assuming all pairs of Cys residues form cystines

Figure 4.22 Analysis of BPSS0238 sequence: Molecular weight, pI, amino acid and atomic composition and extinction coefficient of BPSS0238 were characterised by ProtParam (ExPasy server).

The hydrophobicity of BPSS0238 was analysed using hydropathy plot (http://www.vivo.colostate.edu/molkit/hydropathy/index.html). The Kyte-Doolittle scale plot revealed it is unlikely to be a transmembrane protein (Figure 4.23).

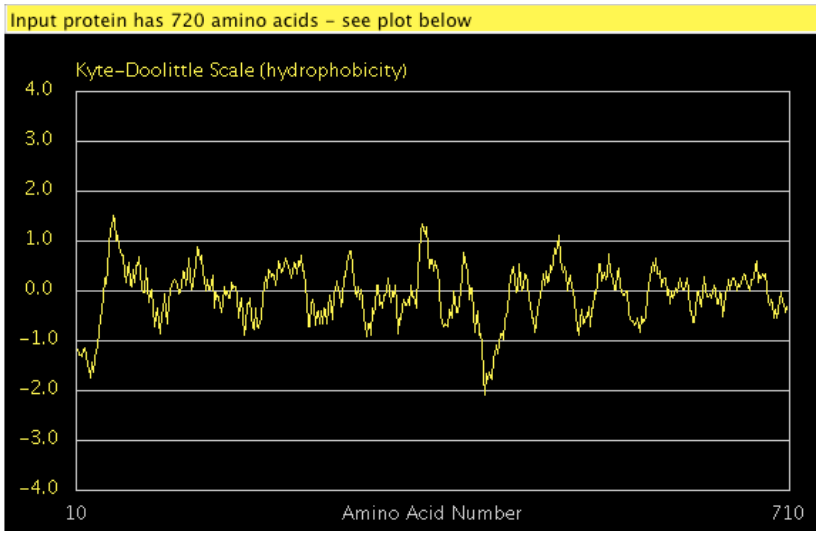


Figure 4.23 A hydropathy plot of BPSS0238: There was no significant feature of transmembrane spanning domain. This figure was produced using the hydropathy plot server (http://www.vivo.colostate.edu/molkit/hydropathy/index.html).

4.5.1.1. Gene amplification and cloning

The oligonuclotide primers (5'-ATGCCAACGAAGGATTTCGAC-3' and 5'-TCAGTCTTCGGTGCGAAACA-3') were used to amplify the gene encoding BPSS0238 by PCR. The PCR product was obtained using an annealing temperature of 56°C with the addition of 10% DMSO.

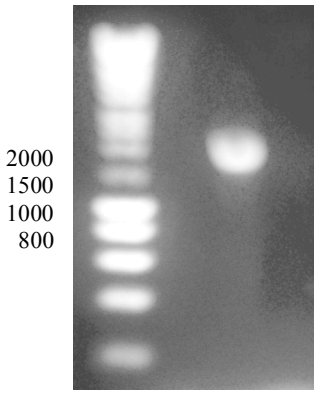


Figure 4.24 PCR product of BPSS0238: The DNA band on a 1% agarose gel corresponds to BPSS0238 with a size of approximately 2000 base pairs.

The PCR product was then cloned using the pETBlue-1 cloning system. Plasmid DNA sequencing of the recombinant clones revealed that the insert was not the BPSS0238 gene. This may be due to the size of this target gene, which is about 2 kb and possibly not suitable for cloning with the pETBlue-1 vector.

Attempts were made to make a new construct using the pET24-a cloning strategy. A new pair of oligonucleotide primers (5'-GTACGCCGCATATGCCAAC-3' and 5'- GGATCCCCGCTCAGTCTTC-3') were used to amplify the BPSS0238 gene. The PCR product was obtained with the expected size of approximately 2000 base pairs and was then cloned into the pET24-a plasmid. The recombinant clone was confirmed by plasmid DNA sequencing. Analysis of the sequencing results showed the identified insert was not the whole target gene of BPSS0238 and was truncated at position 239/240 from the open reading frame of this gene. The BPSS0238 gene contains the sequence G^GATCT, which is similar to the restriction site of *Bam*HI (G^GATCC) that was used in the cloning, resulting in shortening of the insert.

4.5.2. BPSS1416 Uncharacterised protein

The gene encoding BPSS1416, a hypothetical protein in *B. pseudomallei* strain K96243, is located in chromosome 2 and the coding nucleotide sequence consists of 1,404 base pairs encoding a 467-amino acid protein according to the *Burkholderia* genome database (Winsor *et al*, 2008). The protein sequence of BPSS1416 was analysed using ProtParam (ExPasy server), showing its molecular weight of 47.7 kDa with a theoretical pI of 5.5 (Figure 4.25).

```
MPASMSTDQA LRPLLGCIAD DFTGATDLAN MLVKSGMRTV QTIGVPAAGA
AVPADAIVVA LKSRTIPAAD AVAQSLAALD WLRAQGCRQF FFKYCSTFDS
TDAGNIGPVA DALLDALGGE RAFTIACPAF PENGRTVYRG HLFVGDALLG
ESGMENHPLT PMKDANLVRV LQRQTPSKVG LIRHDAIALG TCAVRETIDT
LRREGVRIAI ADALTDRDLY VLGEACADLP LITGGSGVAL GLPSNFRLGG
LLPERGDAAA LPAIGGASAV LAGSASKATH AQVAAWRAER PALRIDPFAA
ARGEPVVDQA LAFARAHLPQ PVLIYASAAP DEVKQVQQAL GIEAAGHLVE
ATLAAIARGL REMGVRKFVV AGGETSGAVV QALGVKALRI GAQIDPGVPA
TATTEGSPCG TTEGSPCGTT EGSPCGATEG SPRGTPDAQP LGLALKSGNF
GSVDFFEKAL RALEGAA
```

Number of amino acids: 467 **Molecular weight:** 47724.5 **Theoretical pI:** 5.51

Amino acid composition:

Ala (A)	82	17.6%	Leu (L)	48	10.3%	Pyl (O)	0	0.0%
Arg (R)	29	6.2%	Lys (K)	11	2.4%	Sec (U)	0	0.0%
Asn (N)	7	1.5%	Met (M)	7	1.5%			
Asp (D)	26	5.6%	Phe (F)	15	3.2%	(B)	0	0.0%
Cys (C)	9	1.9%	Pro (P)	29	6.2%	(Z)	0	0.0%
Gln (Q)	16	3.4%	Ser (S)	22	4.7%	(X)	0	0.0%
Glu (E)	21	4.5%	Thr (T)	29	6.2%			
Gly (G)	52	11.1%	Trp (W)	2	0.4%			
His (H)	6	1.3%	Tyr (Y)	4	0.9%			
Ile (I)	19	4.1%	Val (V)	33	7.1%			

Total number of negatively charged residues (Asp + Glu): 47
Total number of positively charged residues (Arg + Lys): 40

Atomic composition:

Carbon	C	2092
Hydrogen	H	3386
Nitrogen	N	602
Oxygen	O	640
Sulfur	S	16

Formula: $C_{2092}H_{3386}N_{602}O_{640}S_{16}$
Total number of atoms: 6736

Extinction coefficients:

Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

Ext. coefficient 17460
Abs 0.1% (=1 g/l) 0.366, assuming all pairs of Cys residues form cystines

Figure 4.25 Analysis of BPSS1416 sequence: Amino acid composition of BPSS1416 was characterised by ProtParam (Expasy server).

Analysis of the hydrophobicity of BPSS1416 using hydropathy plot in the Kyte-Doolittle scale (http://www.vivo.colostate.edu/molkit/hydropathy/index.html) indicated it is unlikely to be a transmembrane protein (Figure 4.26).

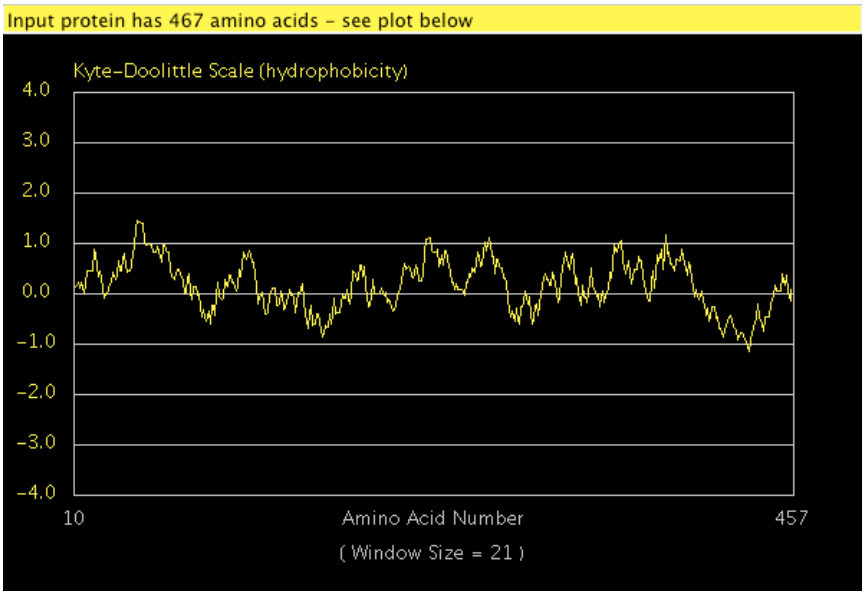


Figure 4.26 A hydropathy plot of BPSS1416: There was no significant hydrophobic region found, and the protein is unlikely to be a membrane spanning protein. This figure was produced using the website (http://www.vivo.colostate.edu/molkit/hydropathy/index.html).

4.5.2.1. Gene amplification and cloning

The gene encoding BPSS1416 was amplified using the oligonucleotide primers (5'- ATGCCTGCCTCCATGAGC-3' and 5'- TCATGCGGCGCCCTCCA-3') by PCR. The PCR product for the target gene fragment was obtained using an annealing temperature of 56°C with the addition of 5% DMSO (Figure 4.27).

The PCR product was then cloned into a pETBlue-1 vector and the recombinant clones were selected using similar procedures before. DNA sequencing of the recombinant plasmids revealed that they were not successfully cloned as the insert was not the target gene. The cloning was repeated but still failed. This may be due to the size of this target gene which is about 1.4 kb and is probably not suitable for cloning with the pETBlue-1 vector.

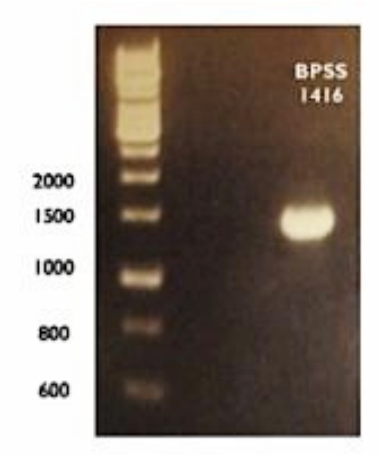


Figure 4.27 PCR product of BPSS1416: The DNA band with approximately 1400 base pairs on a 1% agarose gel corresponds to the gene encoding BPSS1416 (1404 base pairs).

A new construct was thus made of this gene to clone into the pET24-a plasmid using a restriction enzyme cloning strategy. A new pair of oligonucleotide primers (5'-CCGCATATGCCTGCCTCCAT-3' and 5'-AGGATCCTCATGCGGCG CCC-3') were used to amplify the target gene with a number of different PCR conditions but a PCR product that corresponded to the size of the BPSS1416 gene (1404 base pairs) was not obtained.

As it was time-consuming process to achieve the cloning of these two genes, these projects were terminated.

4.6. Other uncharacterised proteins

4.6.1. BPSS0603

The gene encoding BPSS0603 in *B. pseudomallei* strain K96243 is located in chromosome 2 and the coding nucleotide sequence consists of 303 base pairs encoding 100 amino acids according to the *Burkholderia* genome database (Winsor *et al*, 2008).

The protein sequence of BPSS0603 was analysed by ProtParam (ExPASy server). The amino acid composition of BPSS0603 was described, with an approximate molecular weight of 11 kDa, a theoretical pI of 6.55 and an extinction coefficient of 24980 $M^{-1}cm^{-1}$ (Figure 4.28).

```
MIMEASMLNW ISRWALRYAP TPEKSATSML VTARMELFAA EQRVIDAKLQ ADYWCTRVSF
LEVVQKQGID PWVNAQAQKA DDAPAAAVHG AAGPRLAAST
```

Number of amino acids: 100 **Molecular weight:** 10992.6 **Theoretical pI:** 6.55

Amino acid composition:

Ala (A)	20	20.0%	Ile (I)	4	4.0%	Tyr (Y)	2	2.0%
Arg (R)	6	6.0%	Leu (L)	7	7.0%	Val (V)	7	7.0%
Asn (N)	2	2.0%	Lys (K)	4	4.0%	Pyl (O)	0	0.0%
Asp (D)	5	5.0%	Met (M)	5	5.0%	Sec (U)	0	0.0%
Cys (C)	1	1.0%	Phe (F)	2	2.0%			
Gln (Q)	6	6.0%	Pro (P)	5	5.0%	(B)	0	0.0%
Glu (E)	5	5.0%	Ser (S)	6	6.0%	(Z)	0	0.0%
Gly (G)	3	3.0%	Thr (T)	5	5.0%	(X)	0	0.0%
His (H)	1	1.0%	Trp (W)	4	4.0%			

Total number of negatively charged residues (Asp + Glu): 10
Total number of positively charged residues (Arg + Lys): 10

Atomic composition:

Carbon	C	487
Hydrogen	H	768
Nitrogen	N	136
Oxygen	O	142
Sulfur	S	6

Formula: $C_{487}H_{768}N_{136}O_{142}S_6$ **Total number of atoms:** 1539

Extinction coefficients:
Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.
Ext. coefficient 24980
Abs 0.1% (=1 g/l) 2.272, assuming ALL Cys residues appear as half cystines

Figure 4.28 Analysis of BPSS0603 sequence: The characterization of BPSS0603 revealed its molecular weight, theoretical pI, amino acid composition and extinction coefficients by ProtParam.

The hydrophobicity of BPSS0603 was characterised using hydropathy plot, and it can be seen that there were no hydrophobic membrane-spanning regions present in the BPSS0603 sequence (Figure 4.29).

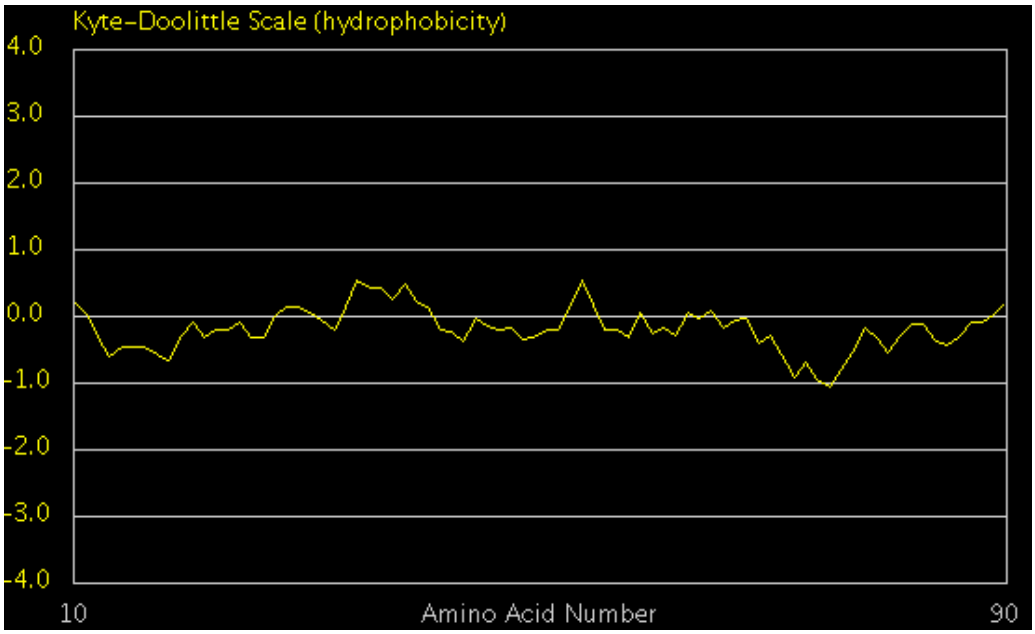


Figure 4.29 A hydropathy plot of BPSS0603, showing that it is unlikely to be a transmembrane protein. This figure was produced using the hydropathy plot server (http://www.vivo.colostate.edu/molkit/hydropathy/index.html).

4.6.1.1. Gene amplification and cloning

The oligonucleotide primers (5'- ATGATAATGGAGGCAAGCATGT-3' and 5'-GTTACGTGCTCGCCGC -3') were used to amplify the target gene using standard PCR. The PCR product was obtained using an annealing temperature of 53°C. The expected size of a 300-base pair band was seen on a 1% agarose gel (Figure 4.30). Due to other non-specific products obtained during PCR, the expected band was cut from the gel and gel extraction was performed to recover the gene fragment prior to cloning.

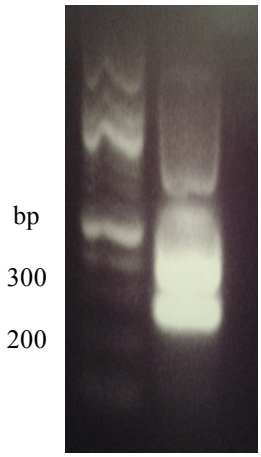


Figure 4.30 PCR product of BPSS0603: The 300 bp band on a 1% agarose gel was cut and the gene fragment was extracted from the gel.

The PCR product was then cloned into a pETBlue-1 vector and the recombinant clones in *E.coli* NovaBlue were selected by the blue/white colony method. White colonies were further screened by colony PCR for the correctly oriented insert clones. The positive clones were confirmed by DNA sequencing. The DNA sequencing results showed that the insert was BPSS0603 but with one amino acid alteration (residue 64 V to E) compared to the reference *B.pseudomallei* strain K96243 genome (Figure 4.31). In comparison to the orthologous proteins from different strains of other completely sequenced genomes of *B. pseudomallei*, the equivalent residues of residue 64 are all valine. The V to E change in this residue may therefore have occurred as a result of a mutation during PCR. Alternatively, it could be a real strain variation. This needs to be clarified.

```
BPSS0603      1 MIMEASMLNWISRWALRYAPTPEKSATSMLVTARMELFAAEQRVIDAKLQADYWCTRVSF
0603_2        1 MIMEASMLNWISRWALRYAPTPEKSATSMLVTARMELFAAEQRVIDAKLQADYWCTRVSF
consensus     1 MIMEASMLNWISRWALRYAPTPEKSATSMLVTARMELFAAEQRVIDAKLQADYWCTRVSF

BPSS0603     61 LEVVQKQGIDPWVNAQAQKADDAPAAAVHGAAGPRLAAST
0603_2       61 LEVEQKQGIDPWVNAQAQKADDAPAAAVHGAAGPRLAAST
consensus    61 LEV-QKQGIDPWVNAQAQKADDAPAAAVHGAAGPRLAAST
```

Figure 4.31 DNA sequencing of the recombinant BPSS0603 clone showed there is one amino acid (colored in red) different from the reference genome (first line).

4.6.1.2. Protein expression

BPSS0603 was over expressed in *E. coli* Tuner™ (DE3) cells with 1mM IPTG added to induce the protein production at 37°C for 4 hours. Analysis on SDS PAGE showed that the target protein (11 kDa) forms inclusion bodies and is in the insoluble fraction (Figure 4.32). An attempt to enhance the solubility of the target protein was carried out with the use of different IPTG concentrations (0.1 mM, 0.25 mM, 0.5 mM and 1mM IPTG) during induction at 37°C, because as the protein synthesis decreases, the protein may be found more in the soluble form. The results showed the protein was still in the insoluble fraction in all conditions. Decreasing the expression temperature to 30°C, still resulted in protein production but it was still insoluble. However, when the over-expression tests were carried out at 20°C, no expression of the target protein was seen.

The alteration of the residue 64 from valine (hydrophobic) to glutamic acid (polar) may have possibly caused the misfolding of the protein, resulting in its presence in the inclusion bodies.

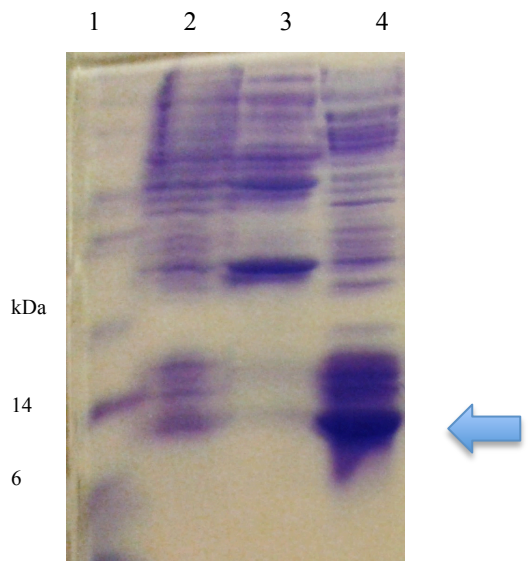


Figure 4.32 Overexpression of BPSS0603: SDS PAGE of BPSS0603 expressed in *E.coli* Tuner™ (DE3) pLacI on a 15% resolving gel; lane 1 a protein marker Mark12™; lane 2 pre-induction; lane 3 soluble fraction at 37°C induction for 4 hours; lane 4 insoluble fraction at 37°C induction for 4 hours.

4.6.2. BPSS0683

According to the *Burkholderia* genome database, the BPSS0683 encoding gene in *B. pseudomallei* strain K96243 is located in chromosome 2 and the coding nucleotide sequence consists of 357 base pairs encoding 118 amino acids (Winsor *et al*, 2008).

The protein sequence of BPSS0683 was analysed by ProtParam (ExPASy server). The amino acid composition of BPSS0683 was described with an approximate molecular weight of 13 kDa, a theoretical pI of 5.09 and an extinction coefficient of 22125 $M^{-1}cm^{-1}$ (Figure 4.33).

```
MRLTIRINGS DAPAQQFAVL WLDTDEQLWS REAHQGIDLP AWGKVKDVEG AVALCSADSG
EALCRLQGLS FSNVRRLSED EEHGNAVLGG KNPQGAWRLQ AVDSASIQPE HREFTIVR
```

Number of amino acids: 118 **Molecular weight:** 12994.5 **Theoretical pI:** 5.09

Amino acid composition:

Ala (A)	13	11.0%	Ile (I)	5	4.2%	Tyr (Y)	0	0.0%
Arg (R)	9	7.6%	Leu (L)	12	10.2%	Val (V)	8	6.8%
Asn (N)	4	3.4%	Lys (K)	3	2.5%	Pyl (O)	0	0.0%
Asp (D)	8	6.8%	Met (M)	1	0.8%	Sec (U)	0	0.0%
Cys (C)	2	1.7%	Phe (F)	3	2.5%			
Gln (Q)	8	6.8%	Pro (P)	4	3.4%	(B)	0	0.0%
Glu (E)	9	7.6%	Ser (S)	9	7.6%	(Z)	0	0.0%
Gly (G)	10	8.5%	Thr (T)	3	2.5%	(X)	0	0.0%
His (H)	3	2.5%	Trp (W)	4	3.4%			

Total number of negatively charged residues (Asp + Glu): 17
Total number of positively charged residues (Arg + Lys): 12

Atomic composition:

Carbon	C	565
Hydrogen	H	892
Nitrogen	N	170
Oxygen	O	177
Sulfur	S	3

Formula: $C_{565}H_{892}N_{170}O_{177}S_3$ **Total number of atoms:** 1807

Extinction coefficients:
Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.
Ext. coefficient 22125
Abs 0.1% (=1 g/l) 1.703, assuming ALL Cys residues appear as half cystines

Figure 4.33 ProtParam of BPSS0683: The protein sequence was analysed and the characterization of BPSS0683 was displayed.

The hydropathicity of BPSS0683 was analysed by hydropathy plot. The hydropathy plot of BPSS0683 revealed that it is unlikely to be a membrane protein (Figure 4.34).

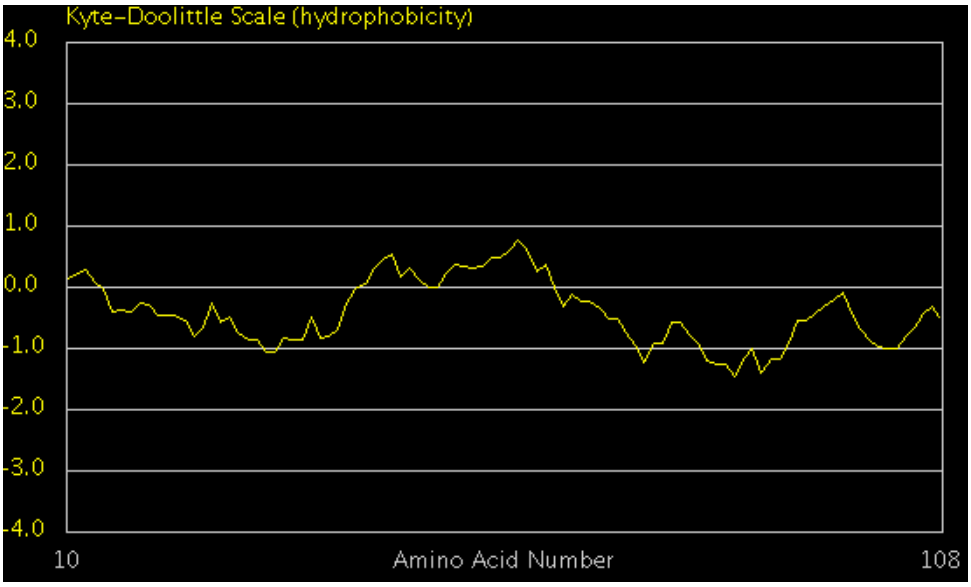


Figure 4.34 A hydropathy plot of BPSS0683 showing that strongly hydrophobic region is not found along the sequence. This figure was produced using the hydropathy plot server (http://www.vivo.colostate.edu/molkit/hydropathy/index.html).

4.6.2.1. Gene amplification and cloning

The gene encoding BPSS0683 was amplified from the genomic DNA of *B.pseudomallei* using the primers 5'-ATGCGCCTCACCATTCGAAT-3' and 5'-TCAGCGCACGATGGTGAAT-3'. The PCR product was obtained using an annealing temperature of 60 °C with an expected PCR product size of about 300 base pairs (Figure 4.35).

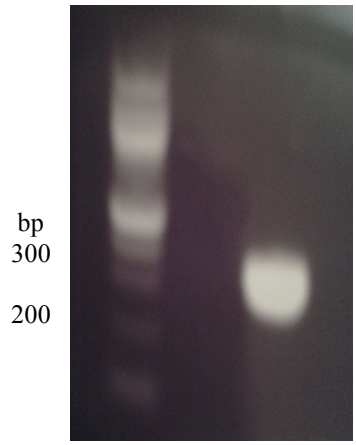


Figure 4.35 PCR product of BPSS0683: An approximately 300 bp band on 1% agarose gel was suspected to be the BPSS0683 gene fragment.

The PCR product was cloned into a pETBlue-1 vector and the recombinant clones were screened by the blue-white colony method and colony PCR. The recombinant clone with the correctly oriented insert was confirmed by DNA sequencing and the results showed that the insert was identical to the reference genome of *B.pseudomallei* strain K96243.

4.6.2.2. Protein expression

The recombinant BPSS0683 was over expressed in *E.coli* Tuner^{TM} (DE3) with 1mM IPTG added to induce the protein production at 37°C for 4 hours but it was found to be insoluble (Figure 4.36). There were no expression of the target protein when the induction temperature was decreased to either 30°C or 20°C.

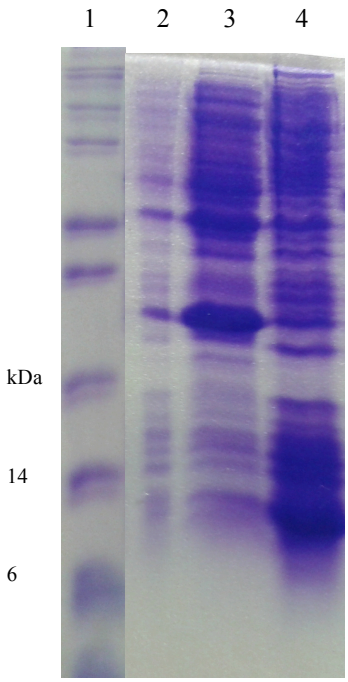


Figure 4.36 Overexpression of BPSS0683: SDS PAGE of BPSS0683 expressed in *E.coli* Tuner^{TM} (DE3) pLacI on a 15% resolving gel; lane 1 a protein marker Mark12^{TM}; lane 2 pre-induction; lane 3 soluble fraction at 37°C induction for 4 hours; lane 4 insoluble fraction at 37°C induction for 4 hours.

4.6.3. BPSL0606

According to the *Burkholderia* genome database, the gene encoding BPSL0606 in the *B. pseudomallei* strain K96243 is located on chromosome 1 and the coding nucleotide sequence consists of 651 base pairs, encoding 216 amino acids (Winsor *et al*, 2008).

A preliminary analysis of BPSL0606 protein sequence was performed using the ProtParam (ExPASy) server. An amino acid composition of BPSL0606 was described with the molecular weight of 23.5 kDa, with a theoretical pI of 6.24 and an extinction coefficient of 20065 $M^{-1}cm^{-1}$ (Figure 4.37).

```
MDWTCCEFRH LSSNELYMIL RTRNAVLVVE DAHTHLDIDG KDEFAIHVFA TDKRGEQPAI
AAYARLLPGD DIDPETTIDK ILTSAAHRDD RTIDALIEHV LAAAHARWPD APVRVQAPAP
REGFYNRFGF RKVDGPYLEH GAPYVGMLRA ASAPSKAVRN LLDLVGNATG ANPGTSAGAV
PASAARKPRK ANAESTTEGE RYAFAGRLPA DSGMNR
```

Number of amino acids: 216 **Molecular weight:** 23486.3 **Theoretical pI:** 6.24

Amino acid composition:

Ala (A) 35	16.2%	Ile (I) 9	4.2%	Tyr (Y) 6	2.8%
Arg (R) 19	8.8%	Leu (L) 16	7.4%	Val (V) 12	5.6%
Asn (N) 8	3.7%	Lys (K) 7	3.2%	Pyl (O) 0	0.0%
Asp (D) 17	7.9%	Met (M) 4	1.9%	Sec (U) 0	0.0%
Cys (C) 2	0.9%	Phe (F) 7	3.2%		
Gln (Q) 2	0.9%	Pro (P) 14	6.5%	(B) 0	0.0%
Glu (E) 12	5.6%	Ser (S) 9	4.2%	(Z) 0	0.0%
Gly (G) 15	6.9%	Thr (T) 12	5.6%	(X) 0	0.0%
His (H) 8	3.7%	Trp (W) 2	0.9%		

Total number of negatively charged residues (Asp + Glu): 29
Total number of positively charged residues (Arg + Lys): 26

Atomic composition:

Carbon	C	1029
Hydrogen	H	1616
Nitrogen	N	308
Oxygen	O	312
Sulfur	S	6

Formula: $C_{1029}H_{1616}N_{308}O_{312}S_{6}$ **Total number of atoms:** 3271

Extinction coefficients:
Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.
Ext. coefficient 20065
Abs 0.1% (=1 g/l) 0.854, assuming ALL Cys residues appear as half cystines

Figure 4.37 The characterization of BPSL0606 sequence by ProtParam (ExPasy server) revealed its amino acid composition, molecular weight, pI and extinction coefficient.

The protein sequence of BPSL0606 was also compared to other proteins in the non-redundant database using BlastP (NCBI). The search showed a hit to the ElaA family protein, which is predicted as a transferase with an acyl-CoA N-acyltransferase domain from *Burkholderia* species and an N-acetyltransferase from other bacteria.

In addition, a domain of the Acyl-CoA N-acetyltransferase superfamily (NAT) has been found at the N-terminal region of the BPSL0606 protein (residues 1-152) using a Superfamily search, the HMM library and a genome assignment server (Gough *et al*, 2001). In comparison to orthologous proteins, BPSL0606 has a C-terminal sequence extension apart from the NAT domain-containing region (Figure 4.38).

Figure 4.38 Schematic of BPSL0606 domain: A NAT domain (residues 1-152) was identified in BPSL0606, with a C-terminal extension not seen in orthologous proteins.

The hydrophobicity of BPSL0606 was characterised using the hydropathy programme (http://www.vivo.colostate.edu/molkit/hydropathy/index.html), which showed that BPSL0606 is not likely to have a transmembrane component (Figure 4.39).

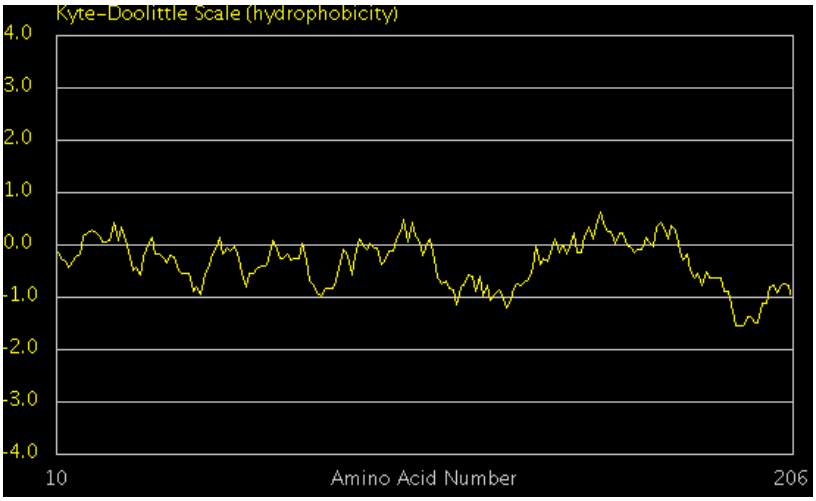


Figure 4.39 A Kyte-Doolittle hydropathy plot of BPSL0606 showed the protein is unlikely to contain a membrane spanning domain. This figure was produced using the hydropathy plot server (http://www.vivo.colostate.edu/molkit/hydropathy/ index.html).

4.6.3.1. Gene amplification and cloning

Oligonucleotide primers were designed for amplifying the full length of the BPSL0606 coding gene according to an open reading frame of this gene in the *B. pseudomallei* strain K96243. The forward primer (5'-GAATCTCCCAT ATGGACTGG-3') and the reverse primer (5'-GGATCCTCATCGGTTCA TTC-3') containing *Nde*I and *Bam*HI restriction sites (underlined) were used in PCR with the genomic DNA of *B. pseudomallei* strain D286.

For amplification of the gene encoding BPSL0606, the PCR product was obtained by adding 2%(v/v) DMSO, resulting in an approximately 600 bp DNA fragment (Figure 4.40).

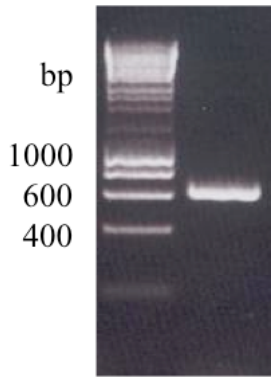


Figure 4.40 Analysis of PCR product of BPSL0606 on a 1% agarose gel electrophoresis: The PCR product has approximate 600 base pairs in length, which corresponds to BPSL0606 (651 base pairs).

The gene fragment of BPSL0606 was inserted into the pET-24a plasmid. Restriction enzyme digestion of the PCR product and the pET-24a vector were carried out to obtain the insert with *Nde*I and *Bam*HI overhangs, and the *Nde*I-*Bam*HI linearised vector, respectively. The insert was then ligated into the vector at the restriction cloning sites using T4 ligase. Recombinant plasmids were transformed into *E. coli* DH5α cells, and screening for positive clones was achieved using the kanamycin selective marker in pET-24a. The transformants were then propagated in a 10 ml culture to isolate plasmid DNA from candidate recombinants for verification.

```
06062f       1 ----------ATTTTGTTTAACTTTAAGAAGGAGATATACATATGGACTGGACATGCTGCGAATTCAGGCATCTGAGTTC
BPSL0606     1 ----------------------------------------ATGGACTGGACATGCTGCGAATTCAGGCATCTGAGTTC
06062r     161 TCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGGACTGGACATGCTGCGAATTCAGGCATCTGAGTTC
consensus  161 ----------attttgtttaactttaagaaggagatatacatATGGACTGGACATGCTGCGAATTCAGGCATCTGAGTTC

06062f      71 GAACGAGCTTTACATGATCCTTCGCACGCGCAATGCAGTCCTCGTCGTCGAGGACGCGCACACGCATCTCGACATCGACG
BPSL0606    39 GAACGAGCTTTACATGATCCTTCGCACGCGCAATGCAGTCCTCGTCGTCGAAGACGCGCACACGCATCTCGACATCGACG
06062r     241 GAACGAGCTTTACATGATCCTTCGCACGCGCAATGCAGTCCTCGTCGTCGAGGACGCGCACACGCATCTCGACATCGACG
consensus  241 GAACGAGCTTTACATGATCCTTCGCACGCGCAATGCAGTCCTCGTCGTCGAgGACGCGCACACGCATCTCGACATCGACG

06062f     151 GCAAGGACGAATTCGCGATCCACGTGTTCGCCACCGACAAGCGCGGCGAGCAGCCGGCGATCGCCGCTTACGCGCGCCTG
BPSL0606   119 GCAAGGACGAATTCGCGATCCACGTGTTCGCCACCGACAAGCGCGGCGAGCAGCCGGCGATCGCCGCTTACGCGCGCCTG
06062r     321 GCAAGGACGAATTCGCGATCCACGTGTTCGCCACCGACAAGCGCGGCGAGCAGCCGGCGATCGCCGCTTACGCGCGCCTG
consensus  321 GCAAGGACGAATTCGCGATCCACGTGTTCGCCACCGACAAGCGCGGCGAGCAGCCGGCGATCGCCGCTTACGCGCGCCTG

06062f     231 CTGCCCGGTGACGACATCGATCCCGAAACGACGATCGACAAGATCCTGACGAGCGCCGCGCACCGCGACGATCGCACGAT
BPSL0606   199 CTGCCCGGTGACGACATCGATCCCGAAACGACGATCGACAAGATCCTGACGAGCGCCGCGCACCGCGACGATCGCACGAT
06062r     401 CTGCCCGGTGACGACATCGATCCCGAAACGACGATCGACAAGATCCTGACGAGCGCCGCGCACCGCGACGATCGCACGAT
consensus  401 CTGCCCGGTGACGACATCGATCCCGAAACGACGATCGACAAGATCCTGACGAGCGCCGCGCACCGCGACGATCGCACGAT

06062f     311 CGACGCGCTGATCGAGCACGTGCTCGCGGCCGCGCACGCACGCTGGCCGGATGCGCCGGTGCGCGTGCAGGCGCCCGCGC
BPSL0606   279 CGACGCGCTGATCGAGCACGTGCTCGCGGCCGCGCACGCACGCTGGCCGGATGCGCCGGTGCGCGTGCAGGCGCCCGCGC
06062r     481 CGACGCGCTGATCGAGCACGTGCTCGCGGCCGCGCACGCACGCTGGCCGGATGCGCCGGTGCGCGTGCAGGCGCCCGCGC
consensus  481 CGACGCGCTGATCGAGCACGTGCTCGCGGCCGCGCACGCACGCTGGCCGGATGCGCCGGTGCGCGTGCAGGCGCCCGCGC

06062f     391 CGCGCGAAGGCTTCTACAACCGCTTCGGCTTCCGCAAGGTCGACGGGCCGTATCTCGAGCATGGCGCGCCGTACGTCGGC
BPSL0606   359 CGCGCGAAGGCTTCTACAACCGCTTCGGCTTCCGCAAGGTCGACGGGCCGTATCTCGAGCATGGCGCGCCGTACGTCGGC
06062r     561 CGCGCGAAGGCTTCTACAACCGCTTCGGCTTCCGCAAGGTCGACGGGCCGTATCTCGAGCATGGCGCGCCGTACGTCGGC
consensus  561 CGCGCGAAGGCTTCTACAACCGCTTCGGCTTCCGCAAGGTCGACGGGCCGTATCTCGAGCATGGCGCGCCGTACGTCGGC

06062f     471 ATGCTGCGCGCGGCGAGCGCCCCGTCGAAGGCCGTGCGCAATCTGCTCGATCTCGTCGGCAACGCGACGGGCGCCAATCC
BPSL0606   439 ATGCTGCGCGCGGCGAGCGCCCCGTCGAAGGCCGTGCGCAATCTGCTCGATCTCGTCGGCAACGCGACGGGCGCCAATCC
06062r     641 ATGCTGCGCGCGGCGAGCGCCCCGTCGAAGGCCGTGCGCAATCTGCTCGATCTCGTCGGCAACGCGACGGGCGCCAATCC
consensus  641 ATGCTGCGCGCGGCGAGCGCCCCGTCGAAGGCCGTGCGCAATCTGCTCGATCTCGTCGGCAACGCGACGGGCGCCAATCC

06062f     551 CGGCACGAGCGCGGGCGCGGTGCCCGCGAGCGCCGCGCGCAAACCGCGCAAGGCCAACGCCGAATCCACCACCGAAGGCG
BPSL0606   519 CGGCACGAGCGCGGGCGCGGTGCCCGCGAGCGCCGCGCGCAAACCGCGCAAGGCCAACGCCGAATCCACCACCGAAGGCG
06062r     721 CGGCACGAGCGCGGGCGCGGTGCCCGCGAGCGCCGCGCGCAAACCGCGCAAGGCCAACGCCGAATCCACCACCGAAGGCG
consensus  721 CGGCACGAGCGCGGGCGCGGTGCCCGCGAGCGCCGCGCGCAAACCGCGCAAGGCCAACGCCGAATCCACCACCGAAGGCG

06062f     631 AACGCTATACGTTCGCCGGCCGTTTGCCGGCCGACTCCGGAATGAACCGATGAGGATCCGAATTCGAGCTCCGTCGACAA
BPSL0606   599 AACGCTATGCGTTCGCCGGCCGTTTGCCGGCCGACTCCGGAATGAACCGATGA---------------------------
06062r     801 AACGCTATACGTTCGCCGGCCGTTTGCCGGCCGACTCCGGAATGAACCGATGAGGATCCGAATTCGAGCTCCGTCGACAA
consensus  801 AACGCTATaCGTTCGCCGGCCGTTTGCCGGCCGACTCCGGAATGAACCGATGAggatccgaattcgagctccgtcgacaa
```

a)

```
BPSL0606     1 MDWTCCEFRHLSSNELYMILRTRNAVLVVEDAHTHLDIDGKDEFAIHVFATDKRGEQPAI
06061/2      1 MDWTCCEFRHLSSNELYMILRTRNAVLVVEDAHTHLDIDGKDEFAIHVFATDKRGEQPAI
consensus    1 MDWTCCEFRHLSSNELYMILRTRNAVLVVEDAHTHLDIDGKDEFAIHVFATDKRGEQPAI

BPSL0606    61 AAYARLLPGDDIDPETTIDKILTSAAHRDDRTIDALIEHVLAAAHARWPDAPVRVQAPAP
06061/2     61 AAYARLLPGDDIDPETTIDKILTSAAHRDDRTIDALIEHVLAAAHARWPDAPVRVQAPAP
consensus   61 AAYARLLPGDDIDPETTIDKILTSAAHRDDRTIDALIEHVLAAAHARWPDAPVRVQAPAP

BPSL0606   121 REGFYNRFGFRKVDGPYLEHGAPYVGMLRAASAPSKAVRNLLDLVGNATGANPGTSAGAV
06061/2    121 REGFYNRFGFRKVDGPYLEHGAPYVGMLRAASAPSKAVRNLLDLVGNATGANPGTSAGAV
consensus  121 REGFYNRFGFRKVDGPYLEHGAPYVGMLRAASAPSKAVRNLLDLVGNATGANPGTSAGAV

BPSL0606   181 PASAARKPRKANAESTTEGERYAFAGRLPADSGMNR
06061/2    181 PASAARKPRKANAESTTEGERYTFAGRLPADSGMNR
consensus  181 PASAARKPRKANAESTTEGERY-FAGRLPADSGMNR
```

b)

Figure 4.41 DNA sequencing of the recombinant BPSL0606 clone: The gene insert in the recombinant pET24-a was identified; a) Nucleotide sequence alignment of BPSL0606 (second line) and positive clone sequenced using T7 forward primer (first line) and T7 reverse primer (third line) shows the alteration of nucleotides in the clone compared to the reference sequence of *B. pseudomallei* strain K96243; b) Protein sequence alignment of BPSL0606 (first line) and the identified insert from positive clone (second line) indicates one amino acid of the clone altered (colored in red) compared to the reference sequence of *B. pseudomallei* strain K96243.

The positive clone was confirmed by DNA sequencing, which showed there are two different nucleotides in comparison to the *B. pseudomallei* strain K96243 genome sequence, resulting in 1 amino acid change at residue 203 (A to T) (Figure 4.41). This sequence difference could be a result of a mutation in the PCR, as this residue in orthologous proteins from different strains of *B. pseudomallei* is always an alanine. Alternatively, it could possibly be a variation between the strains of *B. pseudomallei*. This needs to be clarified.

4.6.3.2. Protein expression

The recombinant protein of BPSL0606 was over produced in *E. coli* BL21 (DE3) cells, with 1 mM IPTG induction at 37°C for 4 hours. The protein expression profile was analysed by SDS PAGE and a large amount of soluble protein was present on the gel between 21.5 and 31 kDa, which was expected as the molecular weight of the target protein is approximately 23.5 kDa. Reducing the temperature to 20 °C during induction did not show a significant increase in the soluble protein yield (Figure 4.42).

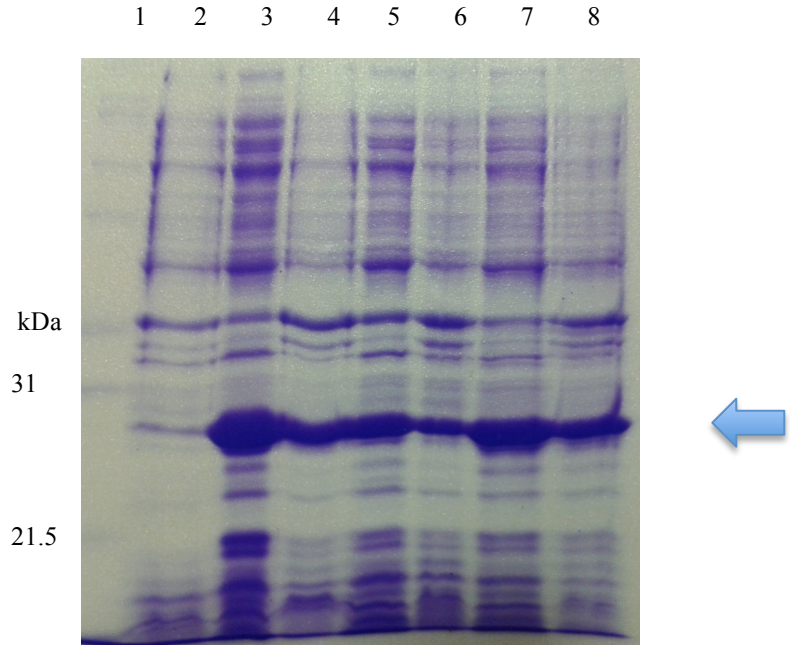


Figure 4.42 BPSL0606 expression profile: lane 1 Mark12TM; lane 2 pre-induction; lane 3 soluble fraction at 37°C induction for 4 hours; lane 4 insoluble fraction at 37°C induction for 4 hours; lane 5 soluble fraction at 20°C induction 4 hours; lane 6 insoluble fraction at 20°C induction 4 hours; lane 7 soluble fraction at 20°C induction overnight; lane 8 insoluble fraction at 20°C induction overnight.

Selenomethionine incorporated protein was also produced for BPSL0606. After growing the bacterial cells to mid-log phase, the use of minimal media with supplements including selenomethionine (described in section 3.3.4) was used in order to allow the production of selenomethionine incorporated protein during induction by IPTG (Stols *et al*, 2004). The supplements added inhibit methionine biosynthesis in the cells and the additional selenomethionine is provided as an external amino acid source for incorporation into the cellular proteins. The overexpression of selenomethionine substituted BPSL0606 was performed using 1 mM IPTG induction at 37°C for 4 hours in a similar manner to the native protein production.

4.6.3.3. Protein purification

Approximately 2 g of cell paste was resuspended in 50 mM Tris-HCl pH 8 and the cells were disrupted by ultra-sonication on ice for 2-3 cycles of 20 seconds. About 20 ml crude extract was obtained after the cell debris was removed by centrifugation. The protein solution was loaded onto a 5 ml Heparin HP column (GE Healthcare) which was pre-equilibrated with 50 mM Tris-HCl pH 8 and the protein was eluted with a gradient of 0-1 M NaCl in 50 mM Tris-HCl pH 8. 2.5 ml fractions were collected and monitored by UV absorbance. The fractions corresponding to the UV peak were analysed using the Bradford assay (Bradford, 1976) and SDS PAGE in order to select the fractions containing the target protein. The second purification step was anion exchange chromatography using a Resource™ Q column (GE Healthcare). The diluted sample was applied to this column, pre-equilibrated with 50 mM Tris-HCl pH 8 and the protein was eluted with a gradient of 0-1 M NaCl in 50 mM Tris-HCl pH 8. 3 ml fractions were collected. The Bio-Rad assay and SDS PAGE were carried out for the fractions corresponding to the UV peak. The fractions containing the purified protein BPSL0606 (23.5 kDa) were pooled together and the protein was concentrated to approximately 10 mg/ml using a 10 kDa MWCO Vivaspin™ centrifugal concentrator (Vivascience).

In the purification steps, BPSL0606, which has its theoretical pI 6.24 predicted by ProtParam (ExPASy server) has a net negative charge in the buffer at pH 8 and binds to the anion exchange column. The purity of BPSL0606 obtained was greater than 95%, which is suitable for subsequent crystallisation trials (Figure 4.43). The average yield of protein purification was approximately 5 mg protein obtained from one litre of recombinant *E. coli* culture.

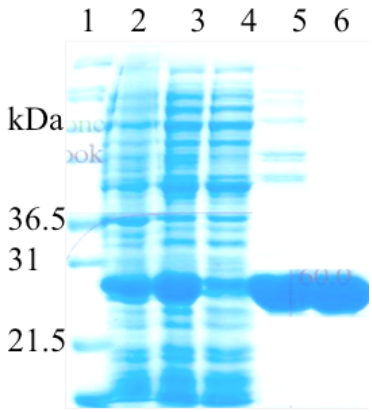
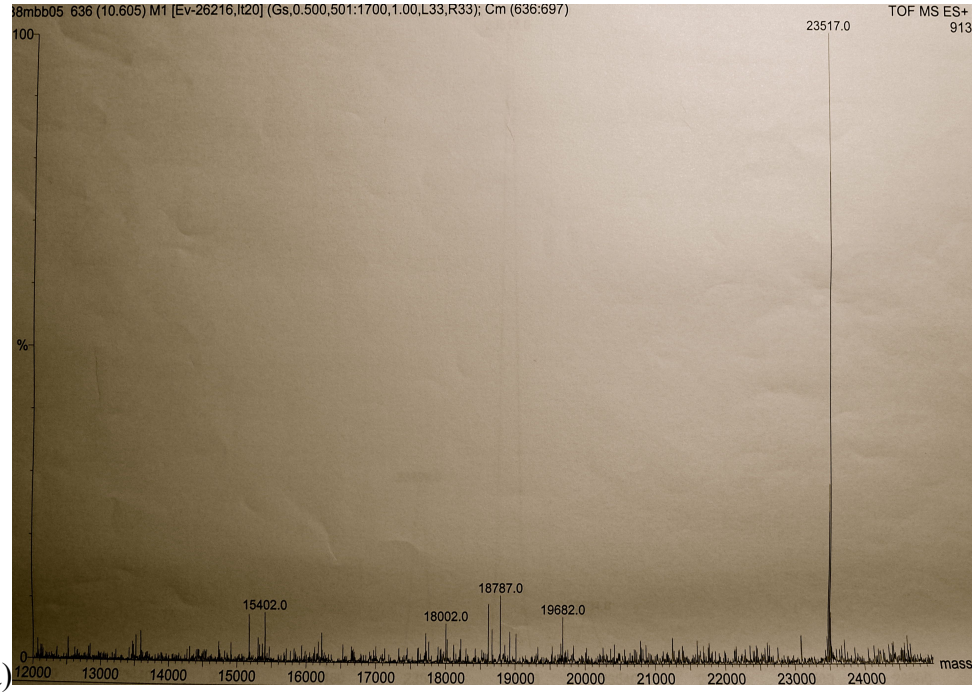


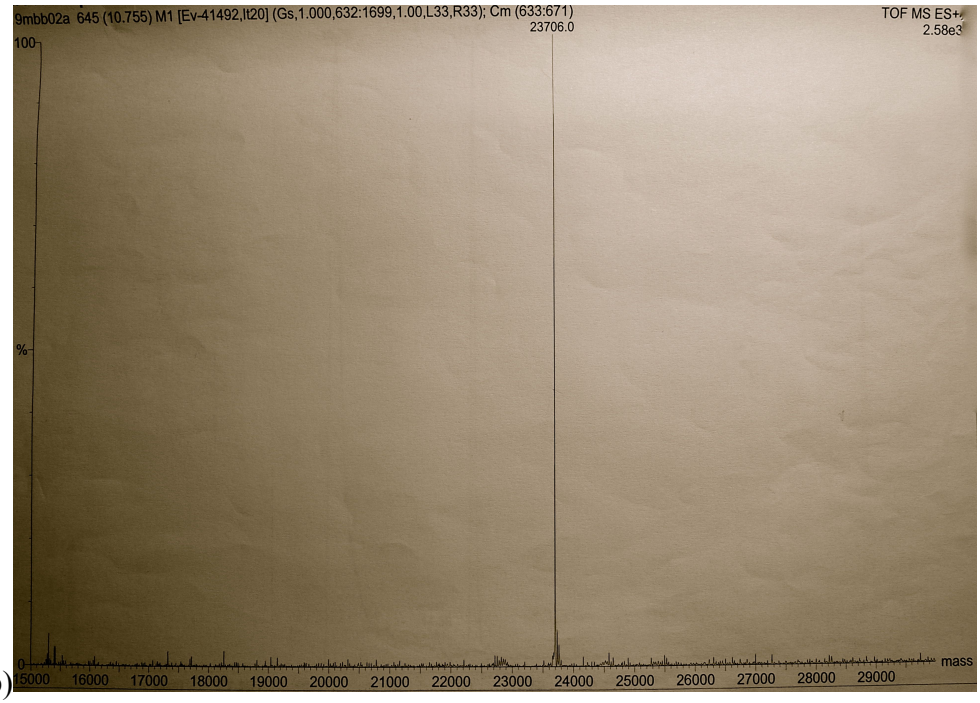
Figure 4.43 Analysis of BPSL0606 protein purification by 12% SDS-PAGE: Lane1 Mark12; Lane2 Cell debris; Lane3 Crude extract; Lane4 Unbound fraction from Heparin HP column; Lane5 Heparin HP column-bound fraction and Lane6 the purified BPSL0606 (23.5 KDa) after Resource Q column.

For the selenomethionine-substituted protein, the same purification protocols were employed. Analysis of molecular weight of the purified protein was carried out by mass spectrometry. The results provided the molecular mass of native BPSL0606 and Se-Met substituted protein (Figure 4.44). The difference between the molecular weights of sulphur-methionine protein and selenium-methionine protein indicate the incorporation of selenium into the methionine residues. As the difference of atomic weights between selenium and sulphur atoms is 78.96-32.06 = 46.9 and BPSL0606 contains 4 methionine residues, the molecular mass of the Se-Met protein would be 187.6 greater than the native protein if selenium is fully incorporated in all 4 methionines. The difference of molecular mass between the Se-Met protein and the native protein obtained from mass spectrometry was calculated to be 189. Thus, BPSL0606 was produced with 100% Se incorporation

of the methionine residues and was thus suitable for subsequent structure determination.



a)



b)

Figure 4.44 Analysis of purified BPSL0606 by mass spectrometry: a) Mass spectrum of native BPSL0606 with its molecular weight of 23517 Da; b) Mass spectrum of Se-Met incorporated protein with its molecular weight of 23706 Da. The difference of molecular weights is approximately the difference between

atomic weights of sulphur and selenium in 4 methionine residues in BPSL0606, indicating the full incorporation of selenium in the Se-Met protein.

4.7. Chapter summary

In summary, from seven selected target genes, five genes including BPSL3022, BPSS0945, BPSS0603, BPSS0683 and BPSL0606 were successfully cloned. Within these five, two proteins, BPSS0945 and BPSL0606, were expressed in a soluble form. These results showed that soluble protein expression remains the bottleneck of protein production, to be used in crystallographic studies. BPSS0945 was unstable and did not crystallise. BPSL0606 was the only protein that was successfully crystallised and whose structure was determined. The crystallisation and structure determination of BPSL0606 is described in the following chapter. The summary of this work is shown in Table 4.5.

Table 4.5 Summary of target proteins studied in this project.

	PCR	Cloning	Sequencing	Overexpression	Purification	Crystallisation
BPSL3022 Short truncated	✓	✓	✓	✗		
BPSL3022 Long truncated	✓ 5% DMSO	✓	1 aa different R→H	✗		
H-tagged periplasmic FtsQ	✓ 2% DMSO	✓	✓	insoluble		
BPSS0238	✓ 10% DMSO	✗				
BPSS1416	✓ 5% DMSO	✗				
BPSS0945	✓ 5% DMSO	✓	2 aa different V→A, A→V	✓	✓	✗
BPSS0603	✓	✓	1 aa different V→E	insoluble		
BPSS0683	✓	✓	✓	insoluble		
BPSL0606	✓ 2% DMSO	✓	1 aa different A→T	✓	✓	*

Possible future experiments for those targets that have not led to structures include:

-BPSL3022: Different constructs of periplasmic FtsQ could be made in order to achieve soluble protein expression and crystallisation. Once the structure of *B. pseudomallei* FtsQ is determined, it could be compared to known structures of FtsQ from other organisms. This could be useful for understanding aspects of cell division in *B. pseudomallei*.

-BPSS0945: Truncated constructs without residues 1-23, which are probably a signal peptide, could be made and the mature protein could be further investigated. Attempts to obtain protein crystals and a structure would provide the structural information and its biological function.

-BPSS0603: A new clone without sequence difference could be made and the soluble protein expression could be observed.

-BPSS0683: To improve the soluble expression of these proteins, co-expression of molecular chaperones such as GroES-GroEL, DnaK-DnaJ-GrpE and ClpB could be helpful for *in vivo* protein folding.

-BPSS0238 and BPSS1416: Different cloning systems could be used for producing the recombinant proteins. As BPSS0238 is a large protein, choosing smaller fragments would perhaps be alternative for cloning this gene and producing the soluble protein.

Sequence differences between the clones and the genomes

All the proteins studied in this thesis have been cloned using genomic DNA from *B. pseudomallei* strain D286. As there was no genome sequence for this strain, the sequences of the successfully cloned genes were compared to the sequences of the genome of *B. pseudomallei* strain K96423, the only *B. pseudomallei* genome sequence available. Problems of assignment thus occur if the sequence of the cloned gene differs from the genome sequence. These differences could either be due to real differences between the strains, or to errors introduced in the PCR process. It is thus important to repeat the cloning of any genes where differences are seen, to identify any potential PCR errors that may result in incorrect folding, solubilisation problems or low yields of the recombinant protein.

For BPSL3022, different sequence changes are seen in the different constructs, with BPSL3022_2 having an A79V change; BPSL3022_3 having one nucleotide difference (no change in protein sequence); BPSL3022_2 having a R175H difference and BPSL3022_1 having the same sequence as that seen in the genome. As the sequence changes are different between these constructs, these differences are thus most likely to be from the PCR reaction.

For BPSS0945, BPSS0603 and BPSL0606, differences are also seen between the cloned gene and the genome sequence. In each of these cases the residue present in the K96423 strain of *B. pseudomallei* in also present in the genes from other *B. pseudomallei* strains, indicating again that an error may have been introduced during the PCR process, alternatively, these sequence differences could be a real strain variation.

For each of these genes the cloning process should be repeated and a number of different clones sequenced, to clarify differences in sequence due to PCR errors from differences due to genetic drift between strains. This may well result in proteins with increased stability, solubility and even an increased chance of crystallisation. However, until a full genome sequence is determined for the D286 strain of *B. pseudomallei*, the cause of these differences in sequence cannot be certain.

Chapter 5: Crystallisation, Analysis of X-ray Diffraction and Structure Determination of BPSL0606

This chapter details the crystallisation of BPSL0606 and the determination of the BPSL0606 crystal structure. Preliminary analysis of the X-ray diffraction data is included.

5.1. Crystallisation of BPSL0606

Freshly purified BPSL0606 protein in 10 mM Tris-HCl pH 8, 0.1 M NaCl was concentrated to approximately 10 mg/ml using a 10 kDa MWCO vivaspin concentrator (Vivascience), prior to crystallisation trials.

Automated crystallisation screening was initially performed using a Hydra II Plus One robot and commercially available screening kits (QIAGEN), including the PACT, JCSG, PEG, pH clear and Classics screens. The experiment was carried out using the sitting-drop vapour-diffusion method, in a 96-well crystallisation plate with each well containing 200 nl of protein mixed with an equal amount of different crystallisation solutions, which consisted of precipitants, buffers and/or salts. All crystallisation trays were sealed using a clear tape and stored at 17°C.

Only a single hit condition from the robot screens was found. BPSL0606 protein crystallised in 20 % (w/v) PEG3350, 0.1 M Bis-Tris propane pH 6.5 and 0.2 M potassium thiocyanate. The crystals of BPSL0606 obtained from the initial robot screening were of a rod morphology, grouped in clusters (Figure 5.1).

To obtain larger crystals, the crystallisation condition for BPSL0606 was further optimised by varying the precipitant concentration, buffer pH and the ratio of protein and reservoir solution using the hanging drop vapour diffusion technique. The optimization of crystallisation included varying the precipitant concentration (14%-24% PEG 3350), the pH (0.1 M Bis-Tris Propane pH 6.3, 6.5 and 7) and the ratio of protein and reservoir solution (1:1, 1:2 and 2:1).

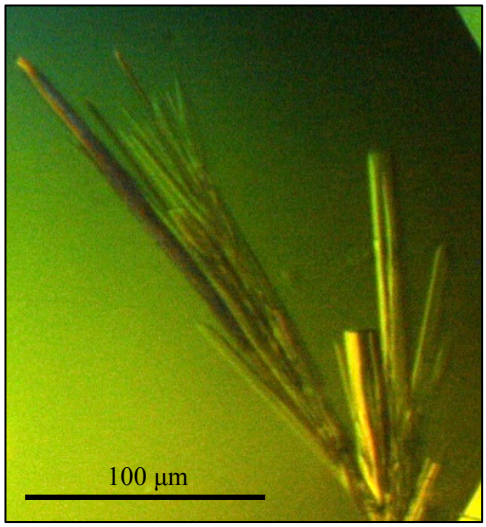


Figure 5.1 Crystals of BPSL0606 observed in the drop F4 in the robot PACT screen. The crystals grew in a solution of 20% (w/v) PEG3350, 0.1 M Bis-Tris propane pH 6.5 and 0.2 M potassium thiocyanate at 270 K.

Single long rod-like crystals were obtained from 16% (w/v) PEG3350, 0.1 M Bis-Tris propane pH 6.3 and 0.2 M potassium thiocyanate, with approximate dimensions of 400 µm x 25 µm x 25 µm.

Prior to mounting on the diffractometer, the protein crystals were mounted in a fibre loop and washed through a cryoprotectant solution, which consisted of the mother liquor they grew from and 25% ethylene glycol. They were then flash cooled to 100 K with an Oxford Cryosystems Cryostream 700 for X-ray diffraction experiments.

The crystals were tested using an 'in house' Rigaku MM007 copper rotating anode generator and a MAR345 Research image plate. Two test images of 1 degree rotation and 5 minutes exposure were taken and the X-ray diffraction patterns were analysed using the autoindexing routine in Mosflm (Leslie & Powell, 2007). The native and Se-M incorporated BPSL0606 protein crystals both belonged to the monoclinic crystal system with the Laue group P2/m and approximate cell dimensions of a=60.6Å, b=81.4Å, c=78.2Å, α=90°, β=100° and γ=90°.

a)

b)

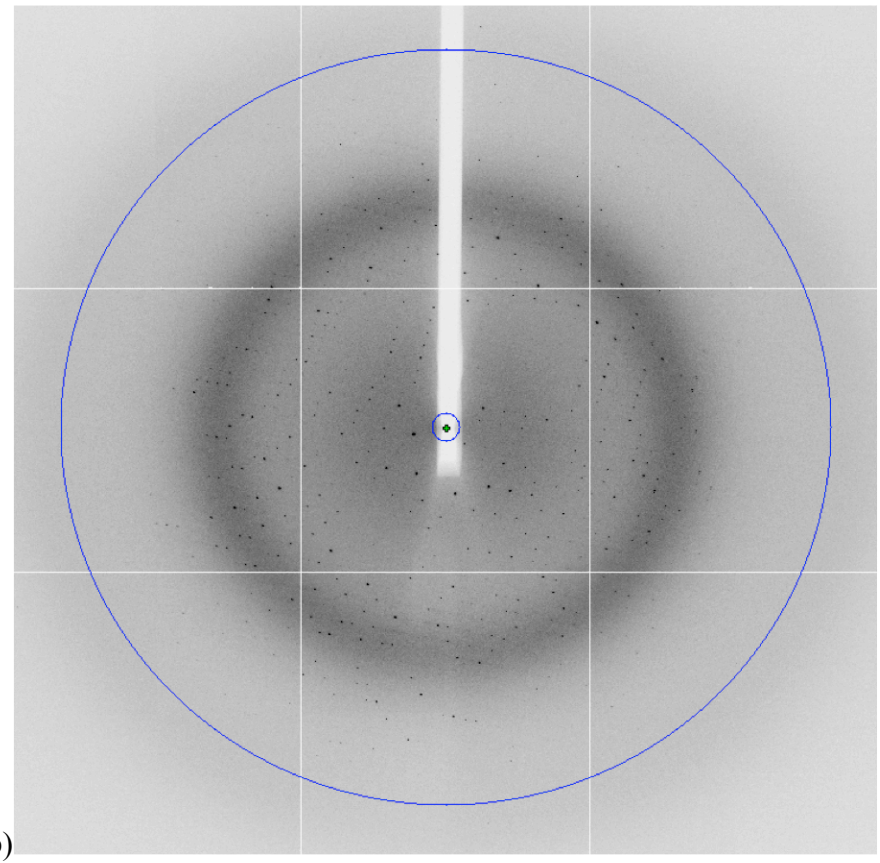


Figure 5.2 Data collection of the native BPSL0606 crystal: a) The sulphur-methionine BPSL0606 crystal was mounted in a fibre loop and was exposed to a 100 µm X-ray beam b) A diffraction image of the BPSL0606 crystal: Diffraction data were collected to a resolution of 2.2 Å on the ADSC Q315 CCD detector at the beamline I03 of the Diamond synchrotron source. The BPSL0606 crystal diffracted the X-rays beyond the resolution of 2.35 Å (blue circle).

5.2. X-ray data collection

Single crystals of BPSL0606 were mounted and cryoprotected with 25% ethylene glycol with their mother liquor prior to cryocooling using an Oxford Cryosystems Cryostream 700 at 100 K. An X-ray diffraction test was carried out using an in-house Rigaku MicroMax 007 micro-focus copper rotating anode generator running at 40kV, 20mA with a MarResearch MAR345 image plate detector system. Each crystal was exposed for 5-15 minutes and a few test images were observed at a starting angle and 90 degree apart to assess the resolution limit and diffraction quality. These images were pre-analysed for its unit cell dimensions, mosaicity and Laue group determination using Mosflm (Leslie & Powell, 2007). Well diffracting crystals were kept in vials in a liquid nitrogen Dewar and they were later transferred to universal pucks for data collection at the Diamond Synchrotron.

High-resolution data sets were collected at the macromolecular crystallography beamlines of the Diamond Synchrotron. DNA software (http://www.dna.ac.uk/) and Mosflm (Leslie & Powell, 2007) were employed for data collection strategies.

The first native data set was collected at beamline I03 of the Diamond Synchrotron source. A total of 360 images with a 0.5° rotation per image were collected to a resolution of 2.2 Å at the edge of the detector.

5.3. Data processing

For the native data set, diffraction images were initially analysed using the autoindexing routine in Mosflm (Leslie & Powell, 2007). Autoindexing was performed to estimate the cell dimensions, crystal system, space group and mosaicity. The possible solution with low penalty indicated the crystal belonged to the monoclinic crystal system in the Laue group P2/m with the cell dimensions of a = 60.61 Å, b = 81.39 Å, c = 78.18 Å, α=90°, β=99.99° and γ=90°, with an estimated mosaicity of 0.63°. The diffraction data were subsequently integrated using Mosflm, resulting in an .mtz file that included the reflection intensities and indices.

To determine the space group, the POINTLESS program (Evans, 2006) in the CCP4 suite was employed. Analysis of systematic absences indicated the presence of 2_1 screw axis along b (0k0; k = 2n), therefore it was likely that the BPSL0606 crystal belonged to space group $P2_1$ (P 1 2_1 1).

Data reduction was performed using the CCP4 package. The data were scaled by the SCALA program (Evans, 2006) and the scaling showed the data were good beyond the resolution of 2.5 Å. Between the resolution of 2.6 Å and 2.4 Å, I/σ fell from 2.3 to 1.4, which is in the range of a typical signal to noise threshold value for a resolution cut-off ($1 < I/\sigma(I) < 2$) (Evans, 2011). The data were therefore cut to a resolution of 2.35 Å.

The unit cell contents were analysed using Matthews_coef program (Kantardjieff & Rupp, 2003; Matthews, 1968) in the CCP4 suite. The unit cell volume was calculated to 3.8×10^5 $Å^3$, resulting in the Matthews coefficient values of 2.71 and 2.03 $Å^3$ Da^{-1} for 3 and 4 molecules in asymmetric unit with 54.6% and 39.5% solvent content, respectively, based on the molecular weight of the full length protein of 23.5 kDa.

Table 5.1 Data collection statistics of the native, sulphur-methonine BPSL0606 crystal

Dataset	
Spacegroup	P 1 2_1 1
Unit cell parameters (Å)	
a	60.61
b	81.39
c	78.18
α	90.00 °
β	99.99 °
γ	90.00 °
Temperature (K)	100
X-ray Source	DIAMOND I03
Detector	ADSC Q315 CCD
Resolution (Å)[1]	29.77-2.35 (2.41-2.35)
Energy (keV)	12.800
Unique observations[1]	30688 (2254)
R_{merge} [1,2]	0.074 (0.583)
R_{pim} [1]	0.055 (0.434)
Completeness (%)[1]	98.4 (98.8)
Multiplicity[1]	3.7 (3.6)
Mean((I)/sd(I))[1]	10.6 (2.2)

[1]Numbers in parentheses indicate values for the highest resolution shell

[2] $R_{merge} = \Sigma_{hkl} \Sigma_i |I_i(hkl) - \langle I(hkl)\rangle| / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where $\langle I(hkl)\rangle$ is the mean intensity of the reflection

5.4. Analysis of self-rotation function and self-Patterson function

A self-rotation function, using the X-ray data between 20-6 Å, shows a peak of 87.3% of the origin at ω = 90.0°, ϕ = 11.7°, κ = 180.0° using an orthogonalization matrix of c along x, a* along y and b* along z. This peak decreases to 85.1% of the origin using 20-3 Å data, and the position changes to ω = 90.0°, ϕ = 10.1°, κ = 180.0°. This gives clear evidence for a non-crystallographic 2-fold axis lying in the ac plane, and approximately 10° away from the c axis (Figure 5.3). Inspection of the self-rotation function showed no significant peaks at κ = 90.0° or κ = 120.0°, indicating that non-crystallographic 4-fold or 3-fold axes are unlikely.

Analysis of self-Patterson using 20-6 Å data shows a peak of 11.5% of the origin at (0, 0.44, 0.47) which decreases to 5.1% of the origin using 20-3 Å data. This perhaps indicates the presence of non-crystallographic translational symmetry.

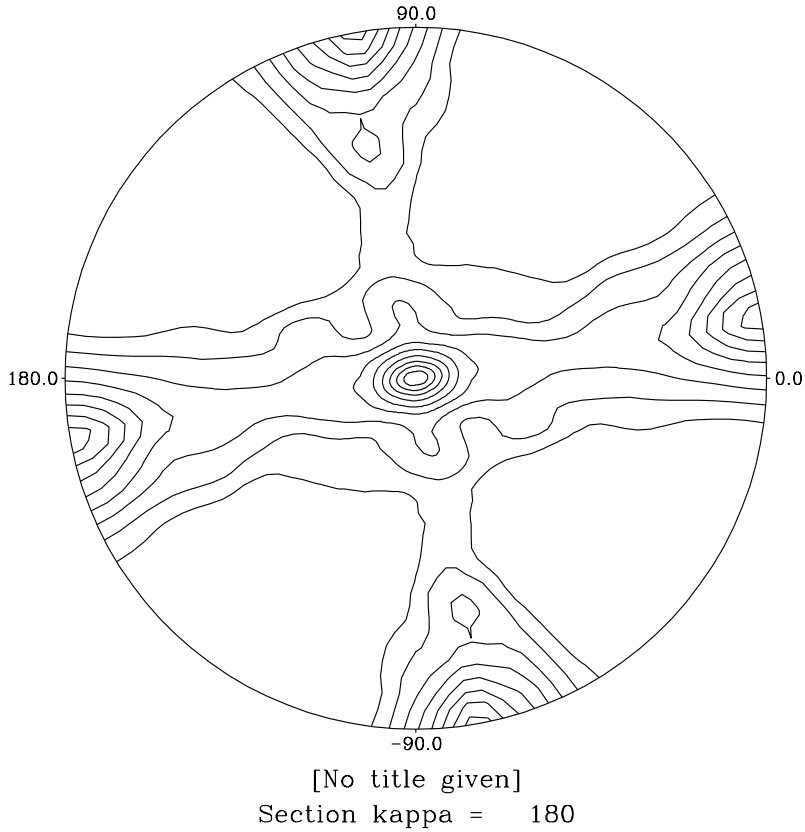


Figure 5.3 Analysis of the self-rotation function using data from 20-6 Å: A peak of 87.3% of the origin can be seen at ω = 90.0°, ϕ = 11.7°, κ = 180.0°, indicating the presence of a non crystallographic 2-fold axis in the ac plane and 11.7° away from the c axis. This figure was generated using Polarrfn in the CCP4 suite (Collaborative Computational Project, 1994).

5.5. Structure determination

As BPSL0606 shares sequence homology with the acyl coA N-acetyltransferase family (Section 4.6.3), then initial attempts were made to determine the structure using molecular replacement. The protein with the highest sequence similarity and known structure to BPSL0606 is an acyl coA N-acetyltransferase from *P. aeruginosa* (PDB ID: 1XEB). These two proteins share 34.9% sequence identity for the N-terminal 150 residues, but the BPSL0606 C-terminal extension of 66 residues is not present in the *P. aeruginosa* protein (Figure 5.4).

Therefore, the coordinates of the crystal structure of the Acyl CoA N-acetyltransferase from *P. aeruginosa* were used as a search model, together with the native BPSL0606 X-ray data for determining BPSL0606 structure by molecular replacement, using the program Phaser (McCoy *et al*, 2007) in the CCP4 package.

```
                    10        20        30        40        50
          CCCCCCEECCCCCHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCEECCCEEC----
1xebA0    SLDWTCKHHADLTLKELYALLQLRTEVFVVEQKCPYQEVDGLDLVGDTHHLMAWRD----
Query     -MDWTCCEFRHLSSNELYMILRTRNAVLVVEDAHTHLDIDGKDEF--AIHVFATDKRGEQ
          -CCEEEEECCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCC--CEEEEEEECCCCC
                    10        20        30        40        50

            60        70        80        90       100       110
          CEECEEEEEECCCCCCCCEEEEEEEECCCCCCCCHHHHCCCHHHHHHHHHHCCCCEEECE
1xebA0    GQLLAYLRLLDPVRHEGQVVIGRVVSSSAARGQGLGHQLMERALQAAERLWLDTPVYLSA
Query     PAIAAYARLLPGDDIDPETTIDKILTSAAHRDDRTIDALIEHVLAAAHARWPDAPVRVQA
          CCEEEEEEEECCCCCCCCEEEEEEEECCCCCCCHHHHHHHHHHHHHHHHHCCCCCEEEEC
           60        70        80        90       100       110

           120       130       140
          ECCCHHHHHCCCCCCCCCCEEECCEEEECCCCC---------------------------
1xebA0    QAHLQAYYGRYGFVAVTEVYLEDDIPHIGMRRA---------------------------
Query     PAPREGFYNRFGFRKVDGPYLEHGAPYVGMLRAASAPSKAVRNLLDLVGNATGANPGTSA
          CCHHHHHHHCCCEECCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHCCCCCCCCCCCCC
          120       130       140       150       160       170


          ----------------------------------------
1xebA0    ----------------------------------------
Query     GAVPASAARKPRKANAESTTEGERYTFAGRLPADSGMNR
          CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
          180       190       200       210

Percentage Identity = 34.9%
```

Figure 5.4 Fold recognition and secondary structure prediction of BPSL0606: The predicted fold recognition and secondary structure of BPSL0606 showed the similarity to acyl coA N-acetyltransferase from *P. aeruginosa* (PDB ID: 1XEB) with 35% sequence identity. This figure was produced using mGenTHREADER (McGuffin *et al*, 2000).

The best solution from the molecular replacement was obtained with a log-likelihood gain (LLG) of 647 and a translation function Z-score of 10.5. This molecular replacement solution contained four copies of the *P. aeruginosa* model each with 149 residues. Although the packing of the four chains in the crystal appeared reasonable (Figure 5.5), electron density maps calculated using this solution did not show any density for the 67 C-terminal residues present in BPSL0606 (Figure 5.6).

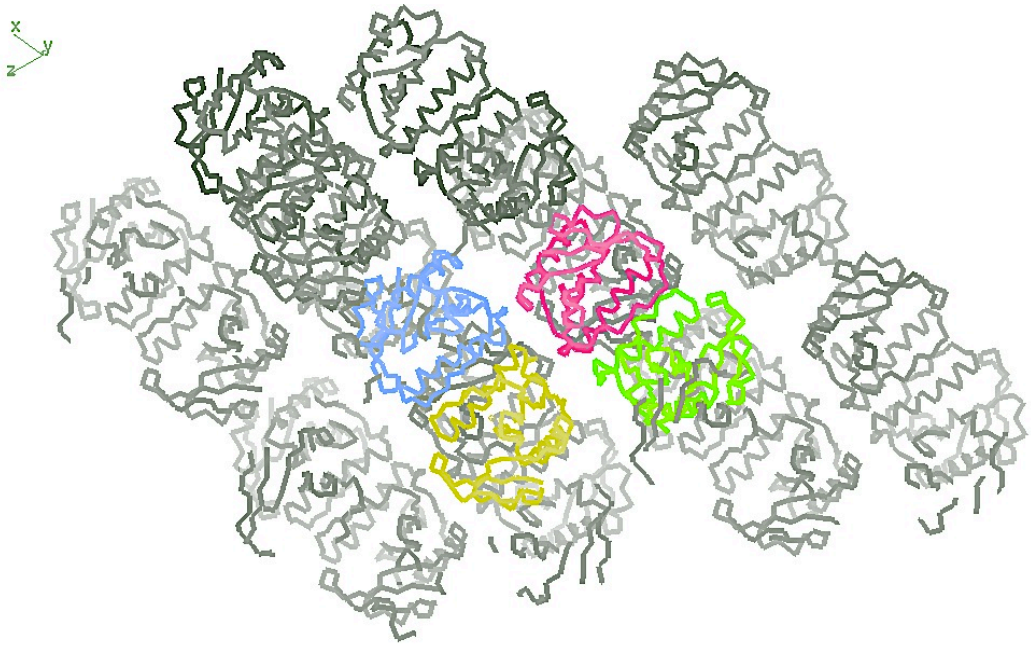


Figure 5.5 Crystal packing for 4 copies of BPSL0606. The 4 molecules in the asymmetric unit pack well in the lattice, and there was no clash between them. Molecules colored in yellow, blue, pink and green represent 4 copies of BPSL0606 and those colored in grey are symmetry related molecules. This figure was produced using Coot (Emsley *et al*, 2010).

Furthermore, the molecular replacement map contained bias against the model, with inconclusive electron density for many of the sequence changes between the 1XEB model and BPSL0606. For example, as can be seen in Figure 5.7, the electron density did not fit the side chain of Ile 93 in BPSL0606 but agreed with the amino acid Gly 93 in the 1XEB search model.

Attempts were made to improve the molecular replacement solution by screening more crystals for better data quality, but no improvement on the quality or resolution of the data could be made.

As the molecular replacement map was not of a very high quality, and because none the C-terminal 65 residues were visible in the map, doubts remained as to whether the molecular replacement result was a correct solution. Therefore, seleno-methionine (Se-Met) incorporated protein was produced to allow independent, unbiased structure determination using the multiwavelength anomalous dispersive (MAD) method.

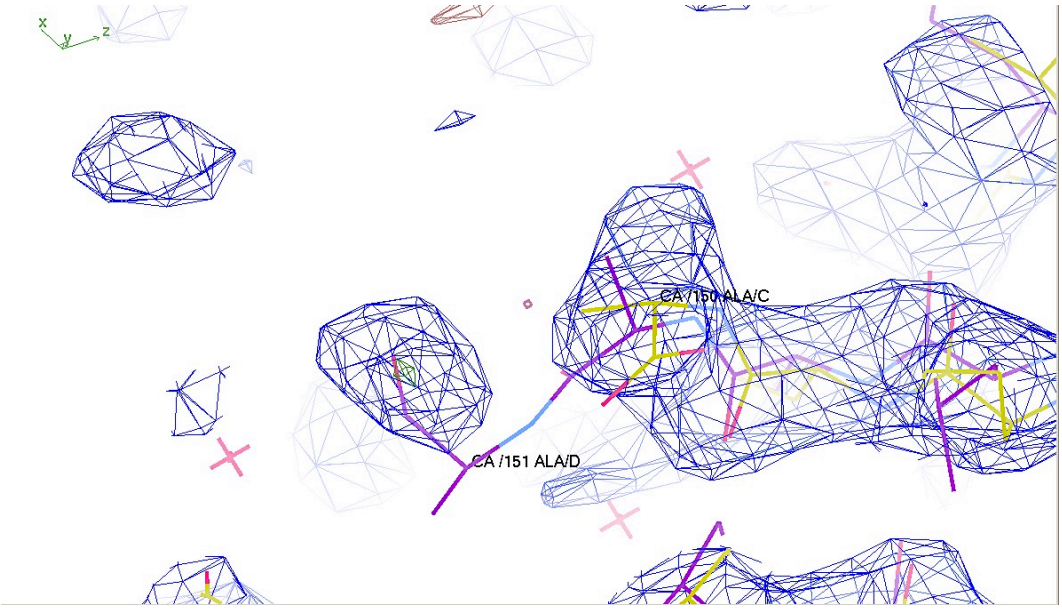


Figure 5.6 The C-terminal 65 residues of BPSL0606 were missing: There was no electron density beyond the residue 151 of BPSL0606. The skeleton of BPSL0606 is shown in violet and the structure of 1XEB, used as a model, is shown in yellow. The electron density map, contoured at 1 σ is shown in blue. This figure was produced using Coot (Emsley *et al*, 2010).

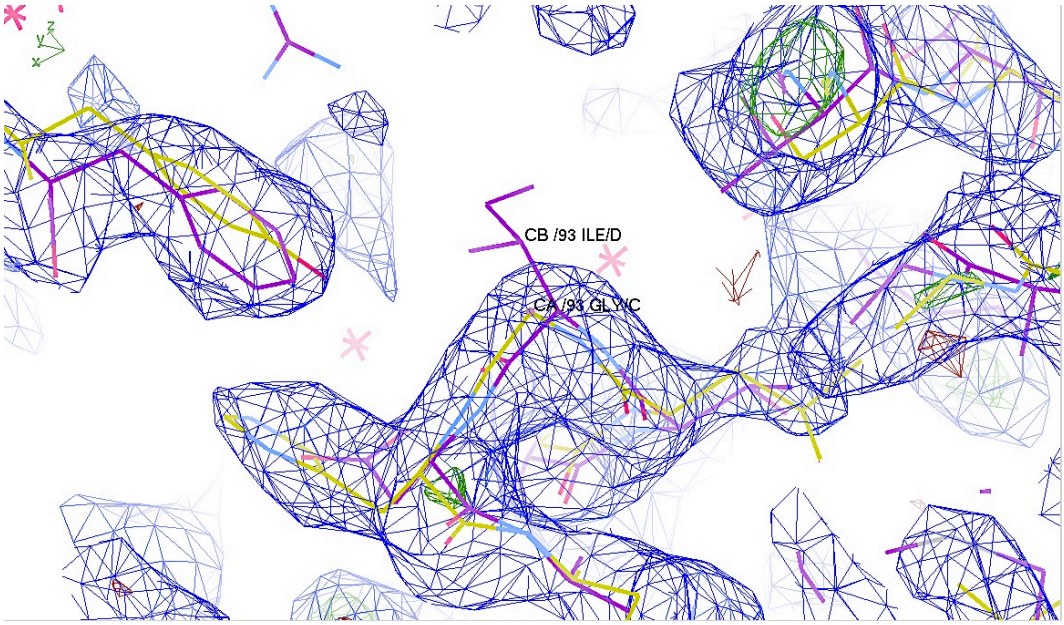


Figure 5.7 Biased structure determination by molecular replacement: The molecular replacement map did not agree in places with the sequence of BPSL0606. The skeleton of BPSL0606 is shown in violet and the structure of 1XEB, used as a model, is shown in yellow. The electron density map, contoured at 1 σ is shown in blue. This figure was produced using Coot (Emsley *et al*, 2010).

5.6. Se-methionine protein

The Se-methionine protein was purified and crystallised in the same conditions as the S-methionine protein, with large rod shaped crystals growing from solutions of 16% (w/v) PEG3350, 0.1 M Bis-Tris propane pH 6.3 and 0.2 M potassium thiocyanate with approximate dimensions of 400 µm x 25 µm x 25 µm (Figure 5.8).

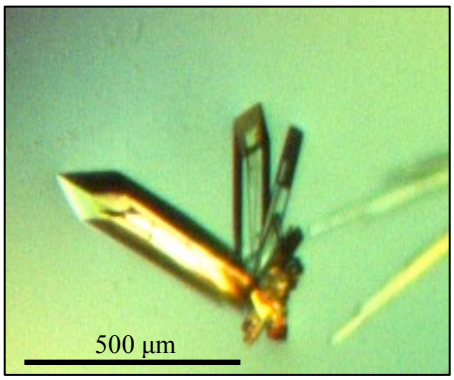


Figure 5.8 Crystals of the Se-Met protein of BPSL0606 grew in solutions of 16% (w/v) PEG3350, 0.1 M Bis-Tris propane pH 6.3 and 0.2 M potassium thiocyanate.

For multi wavelength anomalous dispersion (MAD) experiments, an X-ray fluorescence scan was carried out in order to determine appropriate wavelengths for collecting MAD data using CHOOCH (Evans & Pettifer, 2001).

MAD data were collected at the beamline I04 of the Diamond Synchrotron source. Three wavelengths were selected on a basis of a fluorescence scan at the absorption K-edge of selenium (Figure 5.9). A peak data set was collected at the wavelength λ_1 (λ_1 = 0.9795 equivalent to E = 12657.9 eV) where the f" component is maximised, whereas the inflection data were collected at the wavelength λ_2 (λ_2 = 0.9797 equivalent to E = 12655.4 eV) where the f' component is minimised. The remote data were collected at the wavelength λ_3 (λ_3 = 0.9763 equivalent to E = 12700 eV) in which the dispersive difference is maximised. A total of 180 images with a 1° rotation per image were collected to a resolution of 2.0 Å at the edge of the detector for each wavelength. As the crystals were of a rod morphology, the crystal was translated through the beam while collecting the different data sets to minimise radiation damage (Figure 5.10).

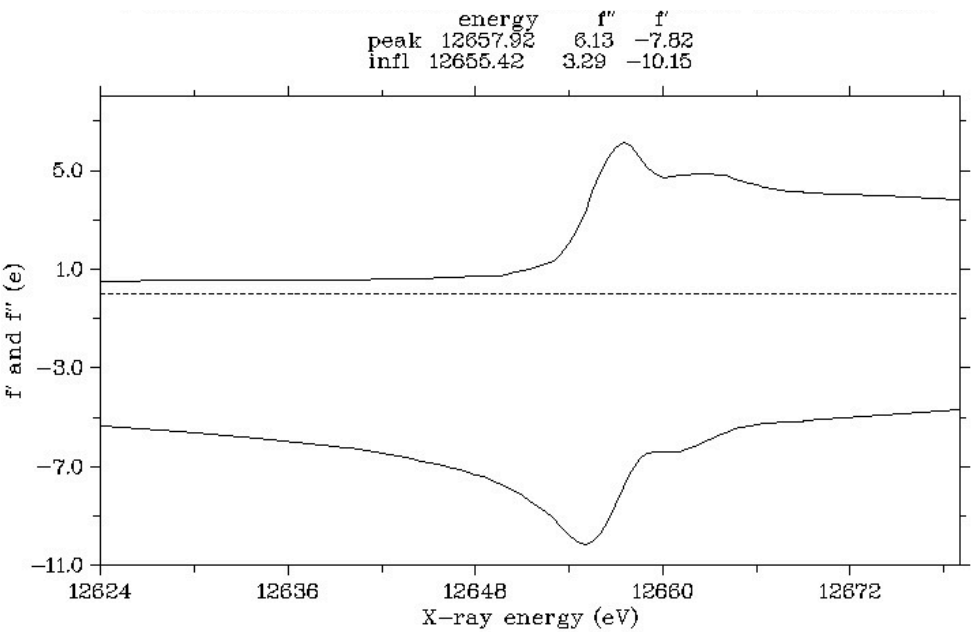
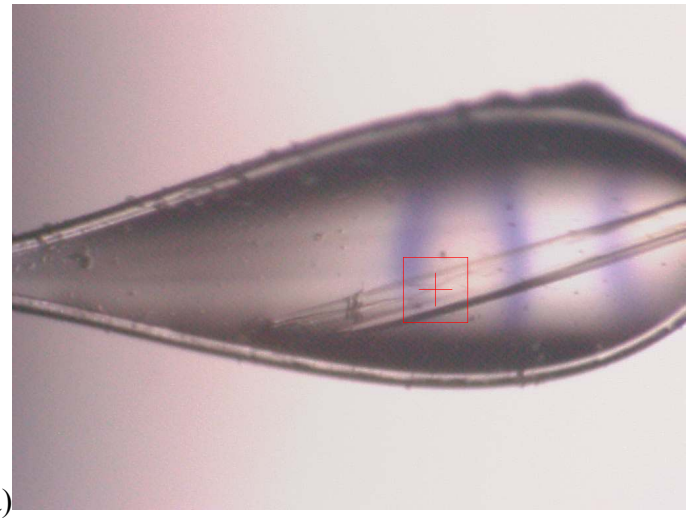


Figure 5.9 A fluorescence absorption spectrum of the Se-Met BPSL0606 crystal near to the absorption K-edge of Se: The absorption signal f" (above) and dispersive signal f' (below) suggested the suitable wavelengths for collecting MAD data sets. This figure was produced using CHOOCH (Evans & Pettifer, 2001).

a)



b)

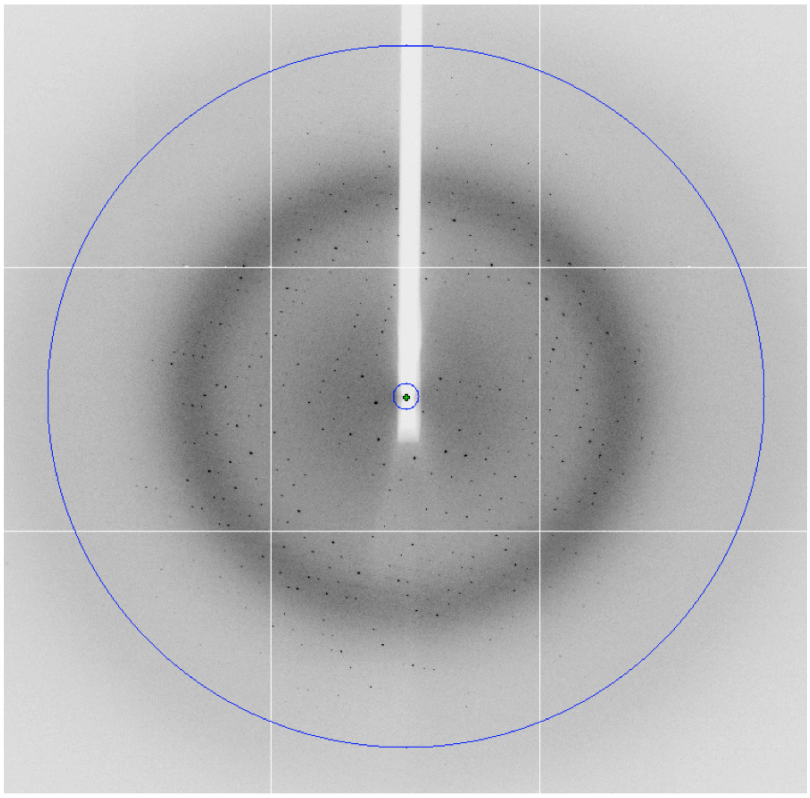


Figure 5.10 MAD data collection: a) A photograph of the BPSL0606 crystal mounted in a fibre loop showed the crystal was being shot by the X-ray beam of 50 µm x 50 µm (red box) and whilst collecting the different data sets, the Se-Met crystal was translated through the beam to minimise radiation damage. b) A diffraction image was collected to a resolution of 2.0 Å at the edge of the ADSC Q315 CCD detector at the beamline I02 of the Diamond synchrotron source. The crystal diffracted the X-rays beyond a resolution of 2.2 Å (blue circle)

For the MAD data, all data sets were indexed, integrated and scaled similarly to the native by the automated program xia2 (Winter, 2010) using the 3dr mode with the XDS and XSCALE programs (Kabsch, 2010), provided at the beamline. Statistics for data collection of the Se-Met protein crystal of BPSL0606 are shown in Table 5.2.

Table 5.2 Data collection statistics of the Se-Met protein crystal of BPSL0606

Datasets	**Se-Met**		
	Inflection	**Remote**	**Peak**
Spacegroup	P 1 2_1 1	P 1 2_1 1	P 1 2_1 1
Unit cell parameters			
a	60.55	60.60	60.62
b	82.10	82.24	82.25
c	78.05	77.98	77.98
α	90.00	90.00	90.00
β	100.14	100.22	100.21
γ	90.00	90.00	90.00
Temperature (K)	100	100	100
X-ray Source	DIAMOND I02	DIAMOND I02	DIAMOND I02
Detector	ADSC Q315 CCD	ADSC Q315 CCD	ADSC Q315 CCD
Resolution (Å)[1]	56.12-2.39 (2.45-2.39)	56.11-2.29 (2.35-2.29)	51.76-2.18 (2.23-2.18)
Energy (keV)	12.6554	12.700	12.6579
Unique observations[1]	29786 (2193)	33910 (2473)	39151 (2829)
R_{merge} [1,2]	0.084 (0.712)	0.089 (0.602)	0.081 (0.639)
R_{pim}[1]	0.065 (0.487)	0.069 (0.42)	0.068 (0.454)
Completeness (%)[1]	99.2 (99.3)	99.2 (98.7)	99.1 (98.6)
Anomalous completeness (%)[1]	95.4 (96.2)	96.0 (95.3)	95.9 (95.3)
Multiplicity[1]	3.7 (3.8)	3.7 (3.8)	3.7 (3.8)
Anomalous multiplicity[1]	1.8 (1.8)	1.9 (1.8)	1.8 (1.8)
Mean((I)/sd(I))[1]	12.5 (1.9)	9.6 (2.2)	10.2 (2.1)

[1]Numbers in parentheses indicate values for the highest resolution shell

[2] $R_{merge} = \Sigma_{hkl} \Sigma_i |I_i(hkl) - \langle I(hkl)\rangle| / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where $\langle I(hkl)\rangle$ is the mean intensity of the reflection

The three-wavelength MAD data sets were input into the automated structure determination program Phenix (Adams *et al*, 2010). The substructure solution from Phenix had 12 selenium sites in total, with a figure of merit 0.43 and overall Z-score 41.6. As the BPSL0606 sequence contains 4 methionines at positions 1, 18, 147 and 214, this substructure is consistent with 3 full-length copies in the asymmetric unit or alternatively, four copies with either the N-terminal methionine or the C-terminal extension missing. Furthermore, Phenix had automatically built 476 residues in 4 chains as an initial model. Clear electron density was visible for the selenium atoms at the position 1, 18 and 147, which accounted for all the peaks in the substructure solution (Figure 5.11). Statistics for all selenium atoms found are shown in Table 5.3.

Table 5.3 Statistics of the selenium sites identified by experimental phasing

Selenium site	**Occupancy**	**B-factor**
Se1	1.21	46.20
Se2	1.07	51.70
Se3	0.88	39.40
Se4	1.04	44.60
Se5	1.1	60.00
Se6	1.01	60.00
Se7	0.69	33.10
Se8	0.77	53.30
Se9	0.45	54.50
Se10	0.50	50.90
Se11	0.59	52.10
Se12	0.56	51.60

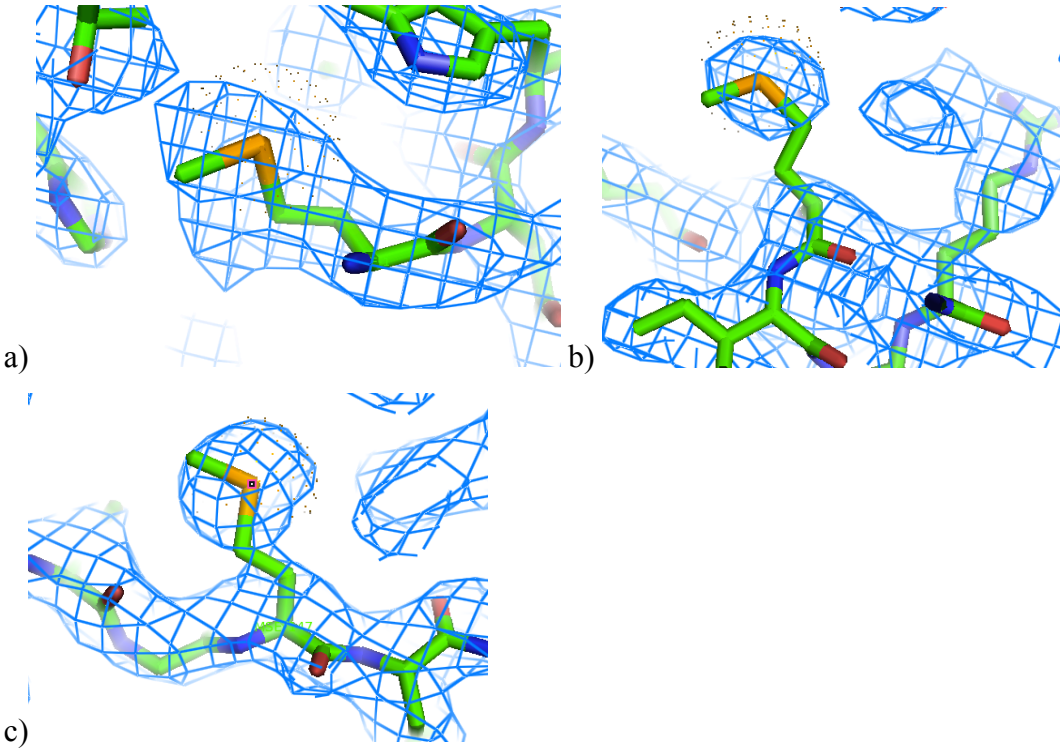
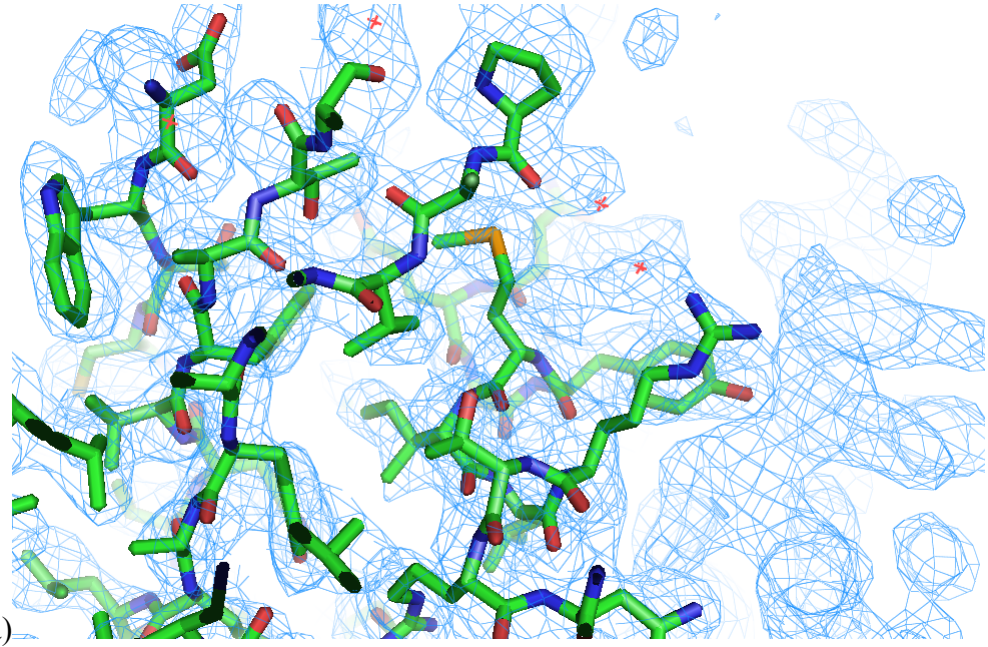


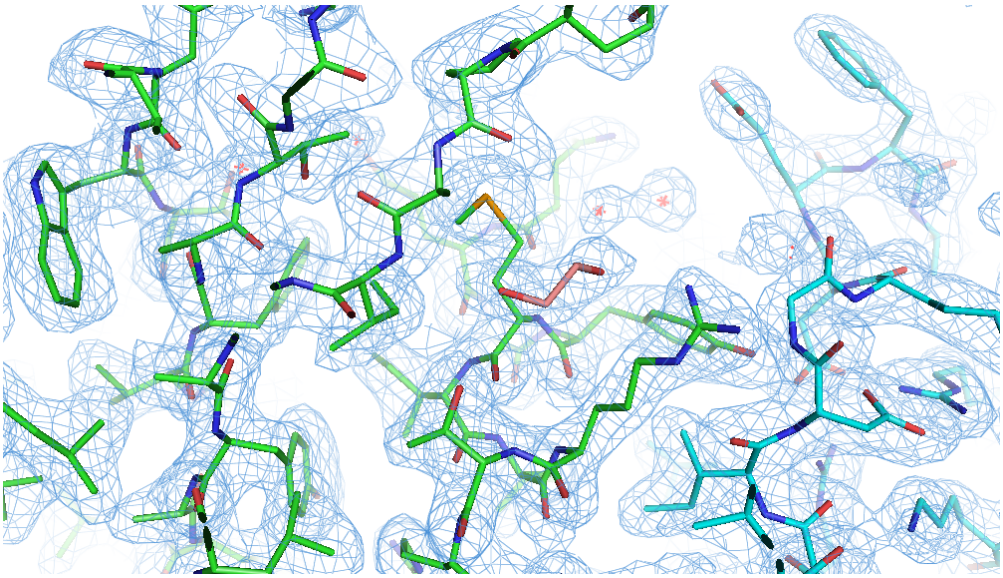
Figure 5.11 Selenium atom location: The initial MAD electron density map, contoured at 1 σ, centered around the selenium atoms in residue 1 (a), 18 (b) and 147 (c). Clear density can be seen to these residues. This figure was produced using Pymol (Schrodinger, 2010).

The initial model was inspected for each residue one by one, to modify residues if appropriate and to join the fragments together. Rebuilding the model to fit the electron density map was performed in COOT (Emsley *et al*, 2010). During the model building task, the model was frequently validated using the validation tools in COOT in order to analyse the geometry, incorrect chiral volumes, peptide omega, rotamers, density fit as well as unidentified blobs and questionable waters throughout the model.

A number of cycles of model refinement and rebuilding were run through the program REFMAC5 in the CCP4 suite (Murshudov *et al*, 2011). Each round of running REFMAC5, the modified model was refined against the peak wavelength data set at 2.18 Å resolution, to avoid bias from any electron density modification from cycle to cycle. The R-factor and R-free were monitored during the refinement tasks (Figure 5.12).



a)



b)

Figure 5.12 A portion of the electron density map of BPSL0606. The model was built and fitted to the electron density map derived from the 2.18 Å resolution data, contoured at 1 σ; a) initial solution, b) final model. This figure was produced using Pymol (Schrodinger, 2010).

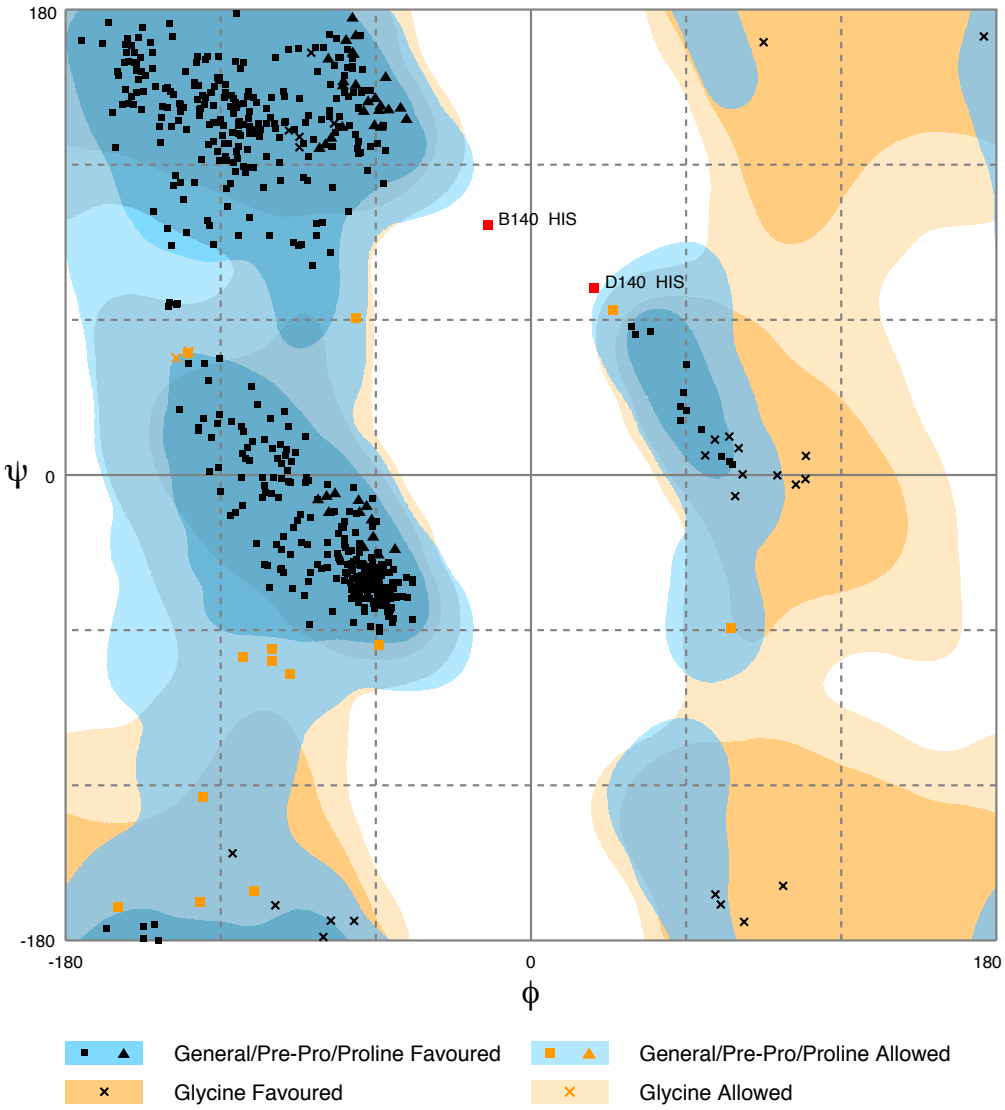
5.7. Structure validation

After a number of refinement cycles were performed, the model of BPSL0606 was analysed using the validation tools in the CCP4 suite such as PROCHECK (Laskowski *et al*, 1993) and RAMPAGE (Lovell *et al*, 2003). The refinement parameters are summarised in Table 5.4. The final model had an R-factor and R-free of 0.196 and 0.25, respectively, which are acceptable at this resolution. Ramachandran plots indicated that 97% of residues were in the favoured region and 0.3% of residues, specifically residues His 140 of chain C and D, were in the outlier region (Figure 5.13). These residues were unfavorable due to poor electron density fit.

Table 5.4 Data refinement statistics of BPSL0606 structure model

Resolution range (Å)	51.76 - 2.18 (2.25 - 2.18)
Space group	P 1 2_1 1
Unit cell dimensions (Å)	a=60.62, b=82.25, c=77.98 α=90° β=100.21° γ=90°
Unique reflections	39131 (3853)
Completeness (%)	98.90 (98.52)
Mean I/sigma(I)	10.19 (2.17)
Wilson B-factor	32.49
Number of molecules in asymmetric unit	4
R-factor	0.1957 (0.2684)
R-free	0.2500 (0.3404)
Number of atoms	5088
macromolecules	4794
water	270
Protein residues	605
RMS(bonds)	0.015
RMS(angles)	1.77
Ramachandran favored (%)	97
Ramachandran outliers (%)	0.34
Clashscore	10.11
Average B-factor	38.00
macromolecules	37.80
solvent	39.20

NB: Statistics for the highest-resolution shell are shown in parentheses.

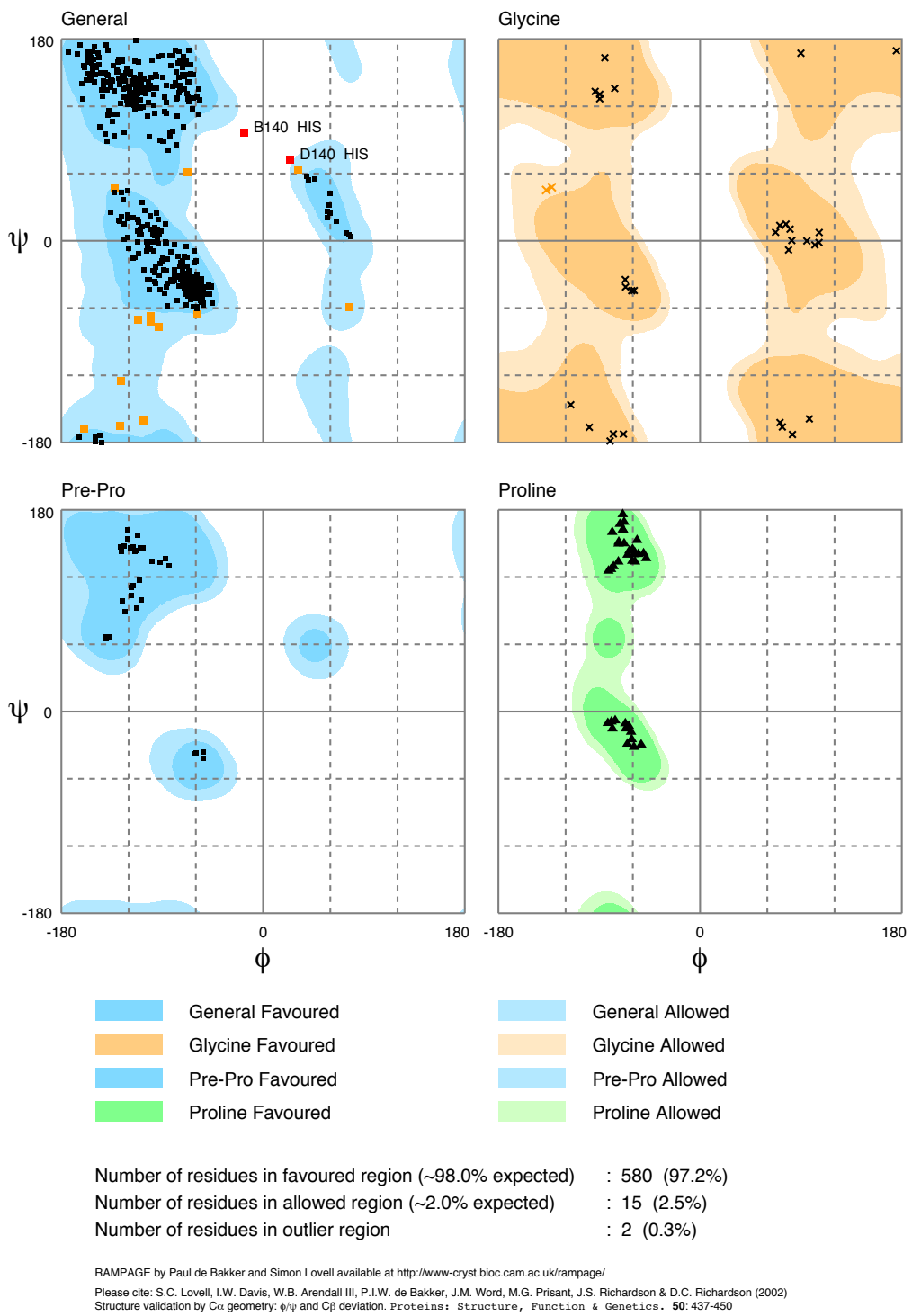


Number of residues in favoured region (~98.0% expected)	: 580 (97.2%)
Number of residues in allowed region (~2.0% expected)	: 15 (2.5%)
Number of residues in outlier region	: 2 (0.3%)

RAMPAGE by Paul de Bakker and Simon Lovell available at http://www-cryst.bioc.cam.ac.uk/rampage/

Please cite: S.C. Lovell, I.W. Davis, W.B. Arendall III, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson & D.C. Richardson (2002)
Structure validation by Cα geometry: ϕ/ψ and Cβ deviation. Proteins: Structure, Function & Genetics. **50**: 437-450

a)



b)

Figure 5.13 Ramachandran plots for the BPSL0606 structure. a) A Ramachandran plot showed the protein backbone torsion angle distribution in general including β-region, right- and left- handed helical regions b) 97% of all residues were in favoured region whereas 2.5% were in the allowed region. Residues His 140 in chain C and D were in the outlier region. This figure was produced using RAMPAGE (Lovell *et al*, 2003).

5.8. Structure description

The final BPSL0606 structure contained 4 copies of the protein, each running from the N-terminal methionine to residue 151 of the full-length sequence. No density was present for the 65 C-terminal residues.

A subunit of the BPSL0606 crystal structure consists of eight β-strands and three α-helices. From the N-terminal region, a short β-strand is connected to the first α-helix (α_1) via a helical loop. Following the long helix, three antiparallel strands (β_2, β_3 and β_4) are formed and precede the second long α-helix (α_2), which is laid underneath the β-sheet. A β-strand (β_5) is linked to the helix α_2 and followed by a short α-helix α_3, turning back behind the strand β_5. Two β strands (β_6 and β_7) are connected and followed by the last strand β_8, parallel to the strand β_5 (Figure 5.14).

In the β-sheets, there is an opening between the strands β_4 and β_5, like a V-cleft. This has the effect of making a groove in structure at the C-terminal ends of the strands β_4 and β_5, which perhaps could be the active site.

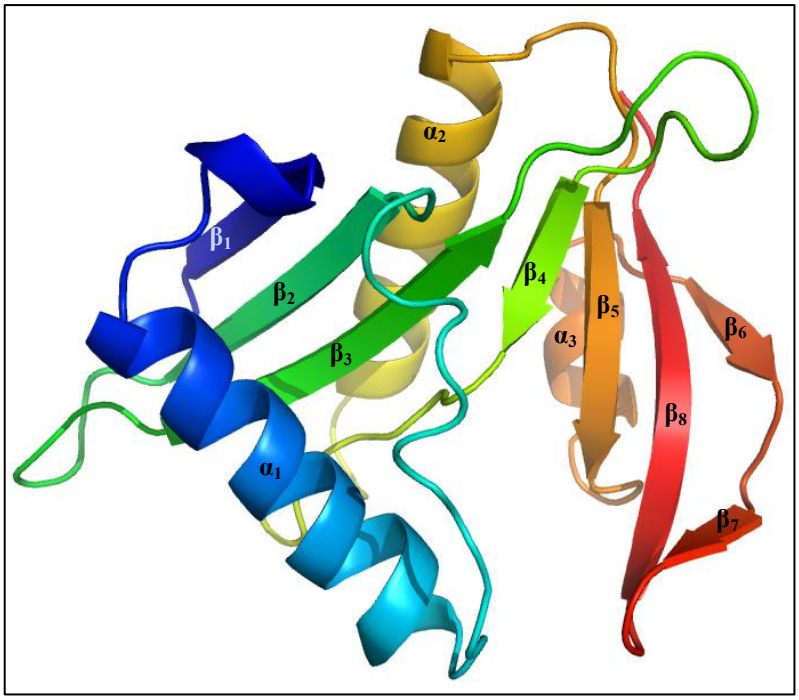


Figure 5.14 The cartoon representation of the secondary structure of BPSL0606 subunit colored from N-terminal (blue) to C-terminal (red). This figure was produced using Pymol (Schrodinger, 2010).

No density was visible for the remaining 65 C-terminal residues, which perhaps indicates that these residues are disordered. Using the RONN server (Yang *et al*, 2005) to predict potential disordered regions of the protein, based on its sequence, there is a high probability of a disordered region from residue 152 to the end, which corresponds to the missing residues in the crystal structure (Figure 5.15).

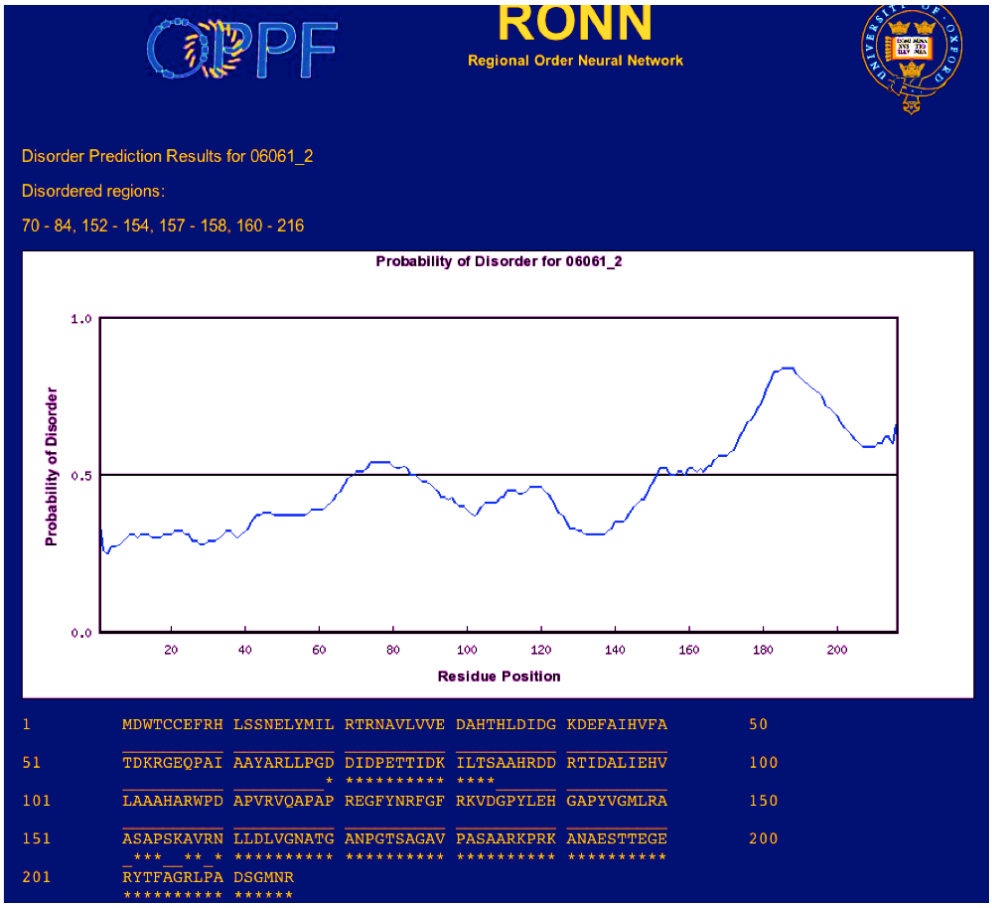


Figure 5.15 Protein disorder prediction: The C-terminal part of the BPSL0606 sequence shows a clear propensity to be a disordered region. This figure was produced using RONN server (Yang *et al*, 2005).

Alternatively, these C-terminal residues may have been cleaved prior to crystallisation. If this were the case, then the 151-residue fragment would have an approximate molecular weight of 17 kDa. Using this value in a cell content analysis, four chains would have a solvent content of 56.3% and Matthews coefficient Vm of 2.81 $Å^3$ Da^{-1}. In comparison, four copies of the full-length protein packed in asymmetric unit would have a Vm of 2.03 $Å^3$ Da^{-1}, corresponding to 39.5% solvent content. It is therefore possible that there is space in the crystal lattice for the remaining 65 C-terminal residues (Figure 5.16).

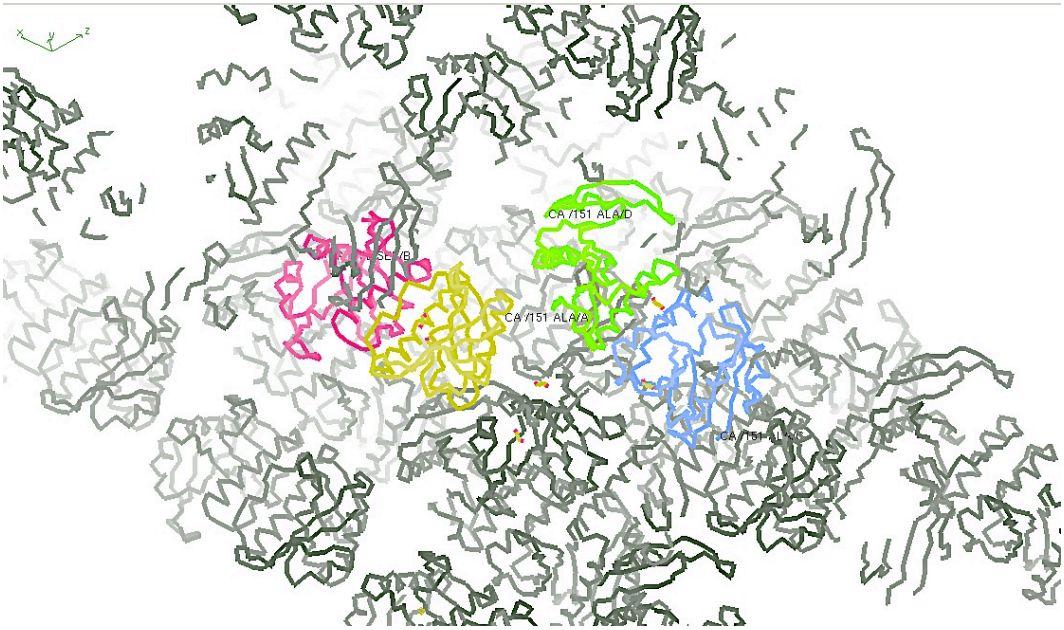


Figure 5.16 Available space in the crystal lattice beyond residue 151: There is room for the missing 65 C-terminal residues. This figure was produced using Coot (Emsley *et al*, 2010).

To investigate whether the crystals were grown from the full-length expressed protein or whether it had been truncated during crystallisation, the remaining crystals in the crystallisation tray were picked out and dissolved in water, prior to analysis using SDS PAGE. A single band from the crystallised protein on the SDS gel lay between 21.5 and 14.4 kDa, indicating that the crystals were not the full-length protein of 23.5 kDa (Figure 5.17). It is therefore likely that after crystallisation was set up with the full-length expressed protein, proteolysis occurred during crystallisation process, resulting cleaved protein being crystallised.

As an analysis of the crystallised protein by mass spectrometry could not be achieved, the same batch of purified protein that had been stored at 4°C for a few months was sent for mass spectrometry instead. The spectrum of the protein mass showed the majority of degraded products were in the range of 17 kDa (Figure 5.18), which corresponds to the obviously intense band of lane 4 on the SDS gel (Figure 5.17). This could explain why the structural model of BPSL0606 was built up to 151 residues instead of the full-length 216 residues.

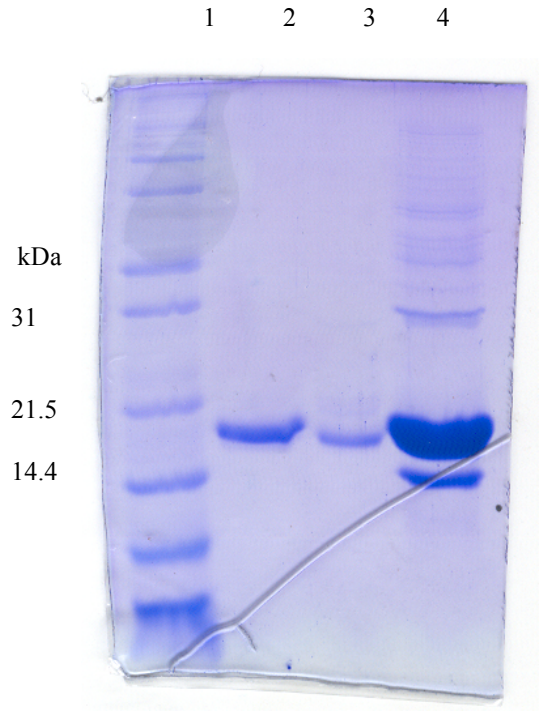


Figure 5.17 SDS PAGE of the BPSL0606 crystals: Lane1 Mark[12] ladder, Lane 2 and 3 crystallised protein, Lane 4 the same purified batch protein as crystallised protein kept in the fridge for a few months.

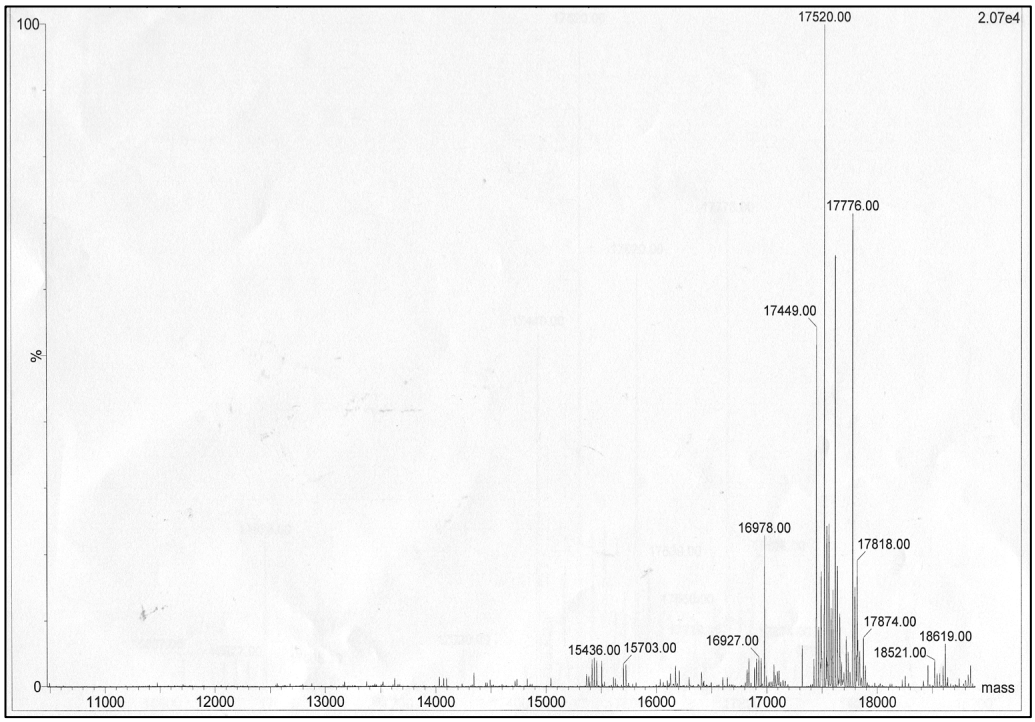


Figure 5.18 Mass spectrometry of the remaining purified protein: From the SDS gel, the remaining protein was degraded from 32 kDa to 17 kDa during a few months storage at 4°C.

Attempts were made to crystallise the remaining protein in the same crystallisation condition that the crystal had grown before but it was not successful. The stored protein was thus analysed further.

5.9. Crystallisation and structure determination of the truncated BPSL0606

Another batch of purified protein BPSL0606 that had been stored in a 4°C refrigerator for 2 months after purification was analysed by mass spectrometry and the spectrum showed the remaining protein was 17.776 kDa, which corresponds to a truncated protein containing 159 residues from N-terminal region (Figure 5.19).

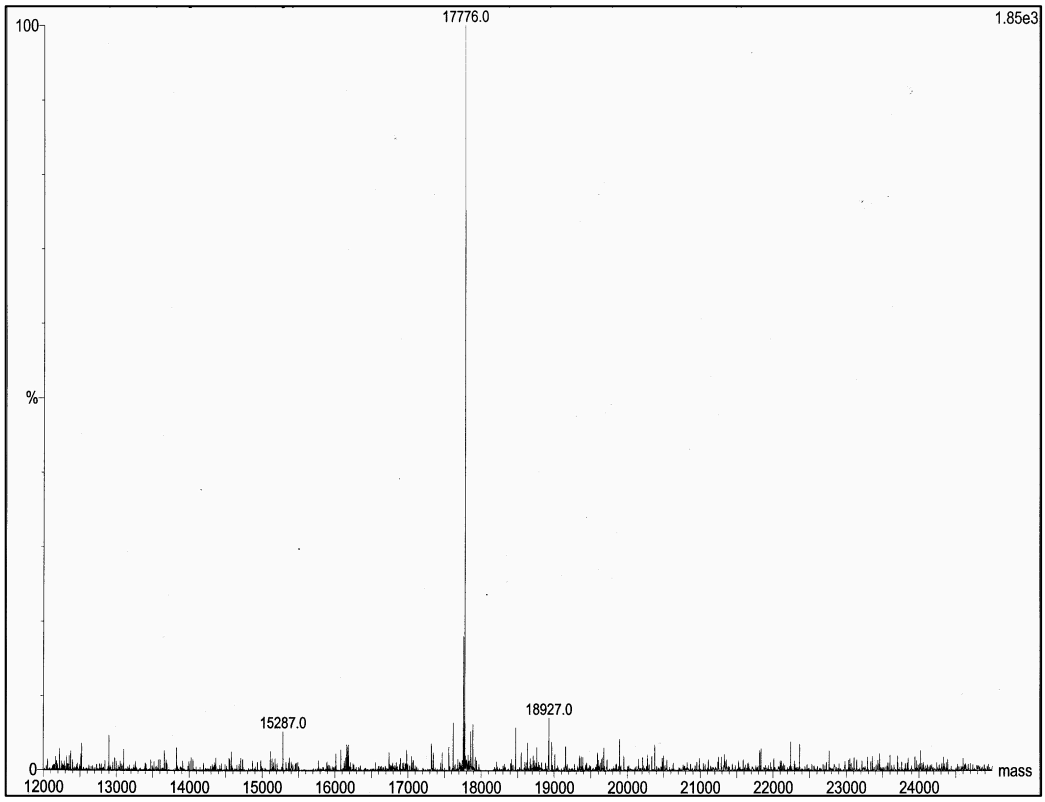


Figure 5.19 Mass spectrometry of the old protein used for the crystallisation experiment: The purified BPSL0606 protein was characterised after keeping at 4 °C for a few months. The spectrum indicated the remaining protein had the molecular weight of 17.8 kDa.

An 8 mg/ml solution of this 17.8 kDa protein was screened for crystallisation by the sitting-drop vapour-diffusion method and the Hydra II Plus One robot with commercially available screening kits (QIAGEN). Approximately 200 µm long single crystals (Figure 5.20) were obtained from the condition H7 in the JCSG screen containing 25% PEG3350, 0.1 M Bis-Tris pH5.5 and 0.2 M ammonium sulphate.

A protein crystal was mounted and washed through a cryoprotectant solution containing 25% ethylene glycol, 25% PEG3350, 0.1 M Bis-Tris pH5.5 and 0.2 M ammonium sulphate, prior to cooling down to 100 K with an Oxford Cryosystems Cryostream 700 for X-ray diffraction tests using an in house Rigaku MM007 copper rotating anode generator and a MAR345 Research image plate. A couple of test images of a few minutes exposure were taken and the X-ray diffraction patterns were analysed using the autoindexing routine in Mosflm (Leslie & Powell, 2007). The truncated BPSL0606 protein crystals belonged to the orthorhombic crystal system, in Laue group P222, with cell dimensions of a=60.94Å, b=75.0 Å, c=78.24Å and α=β= γ=90°.

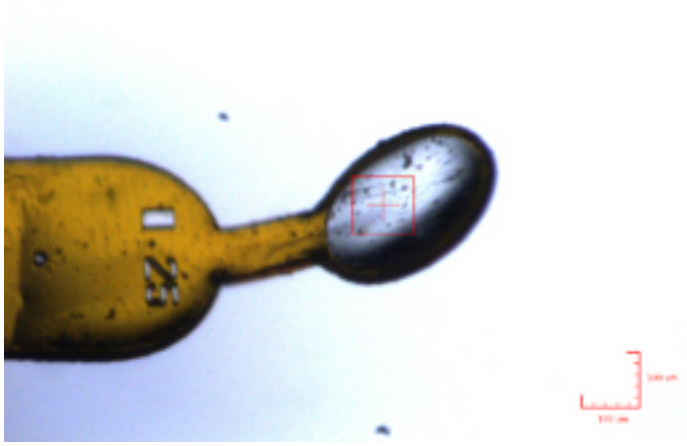


Figure 5.20 A photograph of a cryoprotected crystal of the truncated BPSL0606 mounted by a 0.25 µm loop: The crystal was shot by an X-ray beam of 100 µm x 100 µm (red box) at the Diamond Synchrotron source for data collection.

5.9.1. Data collection and processing

A data set from the truncated protein crystal was collected at the beamline I04 of the Diamond Synchrotron source. A total of 90 images with a 1° rotation per image were collected to a resolution of 2 Å. Statistics of the data collection are shown in Table 5.5.

Table 5.5 Data collection statistics of the truncated BPSL0606 crystal

Data set	
Spacegroup	$P\ 2_1\ 2_1\ 2_1$
Unit cell parameters	
a	60.94
b	75.00
c	78.24
α	90.00
β	90.00
γ	90.00
Temperature (K)	100
X-ray Source	DIAMOND I04
Detector	ADSC Q315 CCD
Resolution (Å)[1]	48.08-2.18 (2.24-2.18)
Energy (keV)	12.649
Unique observations[1]	19172 (1397)
R_{merge} [1,2]	0.089 (0.547)
R_{pim}[1]	0.066 (0.41)
Completeness (%)[1]	99.2 (99.9)
Multiplicity[1]	3.5 (3.6)
Mean((I)/sd(I))[1]	10.3 (2.0)

[1] Numbers in parentheses indicate values for the highest resolution shell

[2] $R_{merge} = \Sigma_{hkl}\ \Sigma_i\ |I_i(hkl) - \langle I(hkl)\rangle| / \Sigma_{hkl}\ \Sigma_i\ I_i(hkl)$, where $\langle I(hkl)\rangle$ is the mean intensity of the reflection

The space group of the truncated BPSL0606 crystal was determined by the POINTLESS program (Evans, 2006). From the analysis of systematic absences, it showed high probability of the presence of the 2_1 screw along all three axes. Therefore, the space group P $2_12_12_1$ was chosen.

Analysis of unit cell content was performed using Matthews_coef program in the CCP4 suite. The unit cell volume was calculated to 3.6 x 10^5 $Å^3$, resulting in the Matthew's coefficient value of 2.51 $Å^3$ Da^{-1} for 2 molecules in asymmetric unit with 51.1 % solvent content for the given the protein molecular weight of 17776 Da.

5.9.2. Structure determination

The structure of the truncated BPSL0606 was determined by molecular replacement using the previously determined BPSL0606 structure. The structure determination was carried out using AutoMR in the automated Phenix program (Adams *et al*, 2010). The best solution had a log likelihood gain (LLG) value of 2,232 and the initial model was automatically built by the AutoBuild program and contained 2 copies of the BPSL0606 protein molecules, 304 residues in total with R-work 0.23 and R-free 0.28.

The model was rebuilt to fit the electron density in COOT (Emsley *et al*, 2010). Several cycles of model refinement against the 2.18 Å data were carried out using REFMAC5 (Murshudov *et al*, 2011). The R-work and R-free of the final model were 0.19 and 0.26, respectively. A summary of refinement statistics is shown in Table 5.6. A portion of the electron density map with the model is shown in Figure 5.21.

In addition to the two protein molecules, a total of 179 water molecules and 9 ethylene glycol molecules could be placed in the electron density map (Figure 5.22).

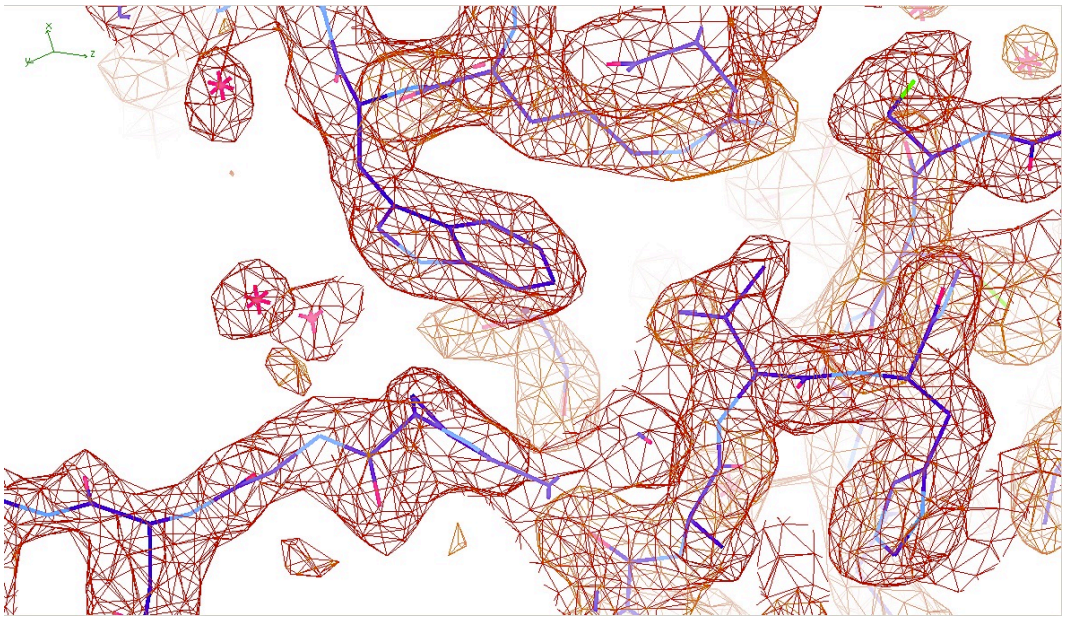


Figure 5.21 The final model of the truncated BPSL0606: The model was refined against the 2.18 Å data. The electron density map, contoured at 1.5 σ, is shown in brown. This figure was produced using Pymol (Schrodinger, 2010).

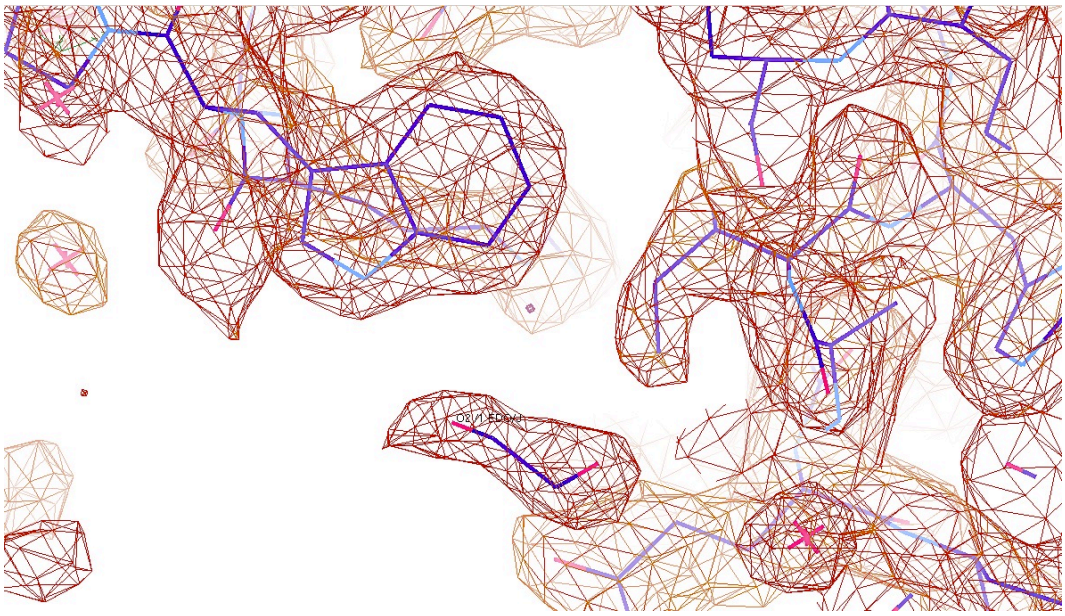


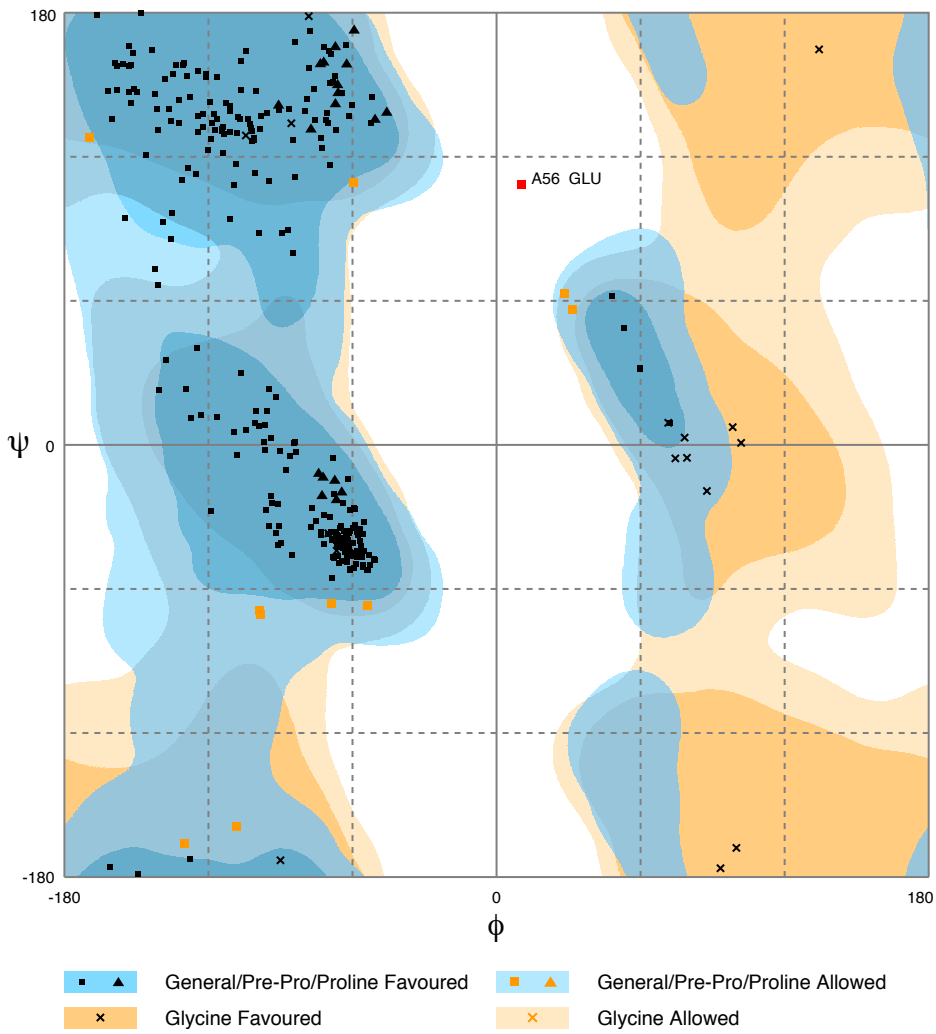
Figure 5.22 Example of other molecules found in the crystal structure: Ethylene glycol used in crystallisation was present as well as water molecules as solvent. The electron density map, contoured at 1.5 σ, is shown in brown. This figure was produced using Pymol (Schrodinger, 2010).

Table 5.6 Data refinement statistics of the truncated BPSL0606

Resolution range (Å)	48.08 - 2.18 (2.258 - 2.18)
Space group	P 2_1 2_1 2_1
Unit cell dimensions (Å)	a=60.94, b=75, c=78.24 α=β=γ=90°
Unique reflections	19127 (1899)
Completeness (%)	99.11 (99.79)
Mean I/sigma(I)	10.37 (2.17)
Wilson B-factor	27.91
R-factor	0.1894 (0.2538)
R-free	0.2593 (0.3463)
Number of atoms	2628
macromolecules	2406
water	179
Protein residues	304
RMS(bonds)	0.014
RMS(angles)	1.71
Ramachandran favored (%)	96
Ramachandran outliers (%)	0.33
Clashscore	6.83
Average B-factor	33.00
macromolecules	32.50
solvent	36.30

Statistics for the highest-resolution shell are shown in parentheses.

Structure validation was performed using PROCHECK (Laskowski *et al*, 1993) and RAMPAGE (Lovell *et al*, 2003) in the CCP4 suite. Ramachandran plots showed that 96% of the total residues were in a favored region and only one residue (Glu 56) in chain A, being an outlier (Figure 5.23). This residue was unfavorable since it has poor electron density for its side chain.

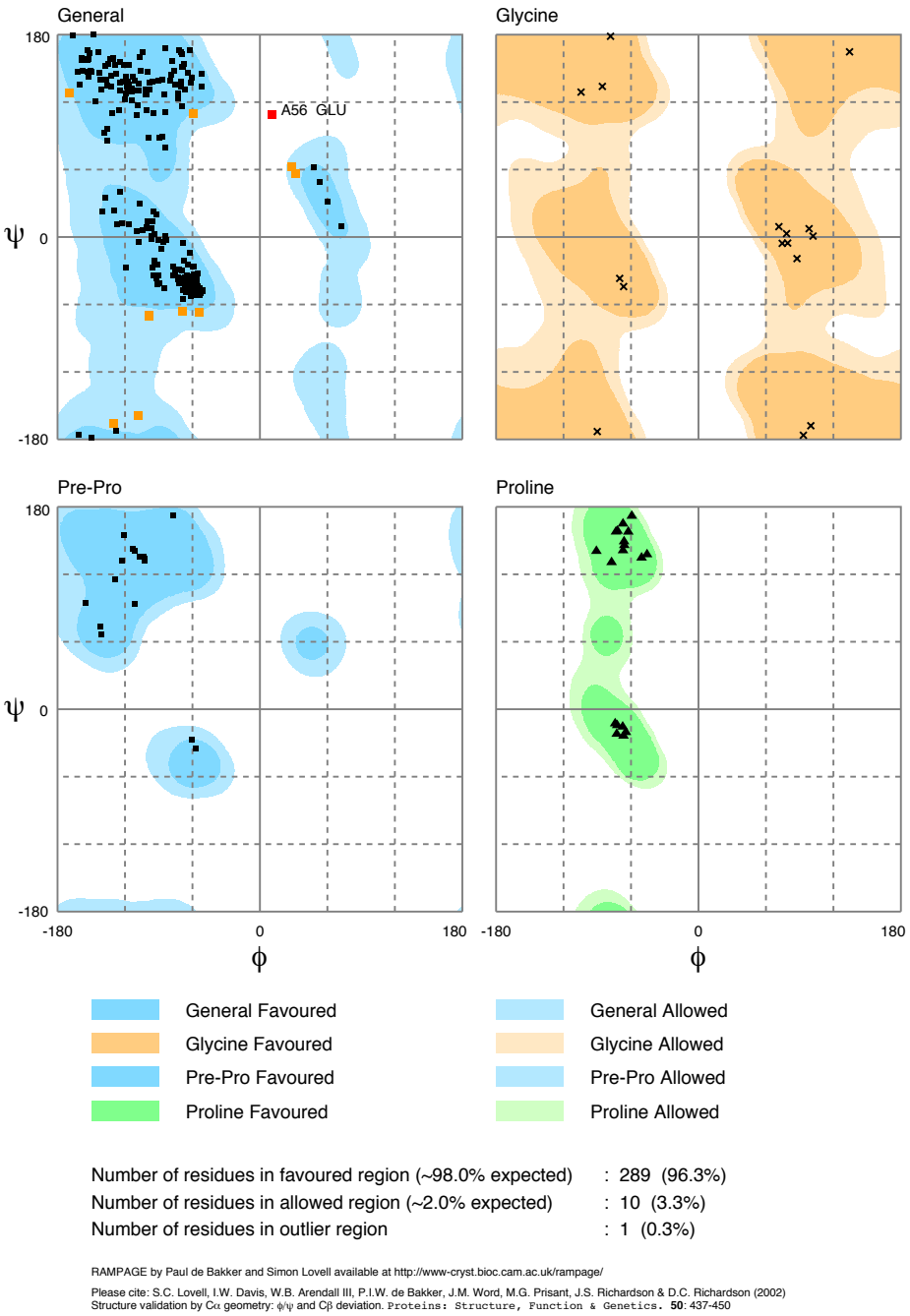


Number of residues in favoured region (~98.0% expected)	: 289 (96.3%)
Number of residues in allowed region (~2.0% expected)	: 10 (3.3%)
Number of residues in outlier region	: 1 (0.3%)

RAMPAGE by Paul de Bakker and Simon Lovell available at http://www-cryst.bioc.cam.ac.uk/rampage/

Please cite: S.C. Lovell, I.W. Davis, W.B. Arendall III, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson & D.C. Richardson (2002) Structure validation by Cα geometry: ϕ/ψ and Cβ deviation. Proteins: Structure, Function & Genetics. **50**: 437-450

a)



b)

Figure 5.23 Ramachandran plots for the truncated BPSL0606 structure: a) A Ramachandran plot showed the protein backbone torsion angle distribution in general including β-region, right- and left- handed helical regions b) 96% of all residues were in favoured region whereas 3.3% were in the allowed region. Residue Glu 56 in chain A was in the outlier region. This figure was produced using RAMPAGE (Lovell *et al*, 2003).

5.9.3. Structure description

The topology of the truncated BPSL0606 structure was exactly the same as the previously determined structure of BPSL0606. The conformation of the truncated BPSL0606 structure was similar to the BPSL0606 subunit that had been determined before. In the crystal structure of the truncated BPSL0606, there were two molecules in asymmetric unit and each molecule contained 152 residues.

The five residues 153-159, that from the mass spectrometry results were presumably present in protein being crystallised were again not visible in the map, showing that these residues are either disordered or a further truncation has occurred during crystallisation.

Unlike the previously determined BPSL0606 structure, the crystal structure of the truncated BPSL0606 was different in space group and number of molecules in asymmetric unit. The truncated BPSL0606 crystal belonged to the space group $P2_12_12_1$ and consisted of two copies of molecule in the asymmetric unit with approximate 51% solvent content whereas the BPSL0606 crystal was in $P2_1$, filled with four molecules in asymmetric unit with 40% solvent content.

Analysis of the crystal structures of BPSL0606 and structure comparison will be detailed in the next chapters.

Chapter 6: Structure of BPSL0606

This chapter details the crystal structures of BPSL0606. Analysis of the individual protein subunits and the protein quaternary structure is included.

6.1. The crystal structure of BPSL0606 contained 4 molecules in the asymmetric unit ($P2_1$ crystal form)

The full-length BPSL0606 protein was put into crystallisation trials and crystals containing truncated protein were obtained. The structure built from the electron density that was derived from the X-ray diffraction pattern consisted of 151 residues of each molecule of 4 copies found in the asymmetric unit. The crystal structure of BPSL0606 is shown in Figure 6.1.

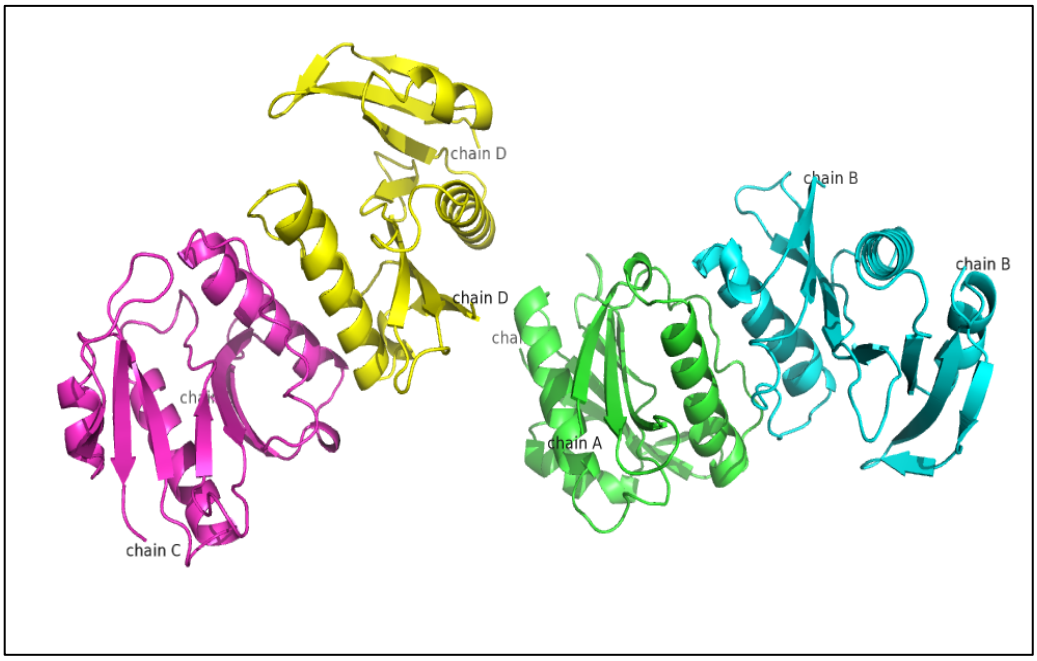


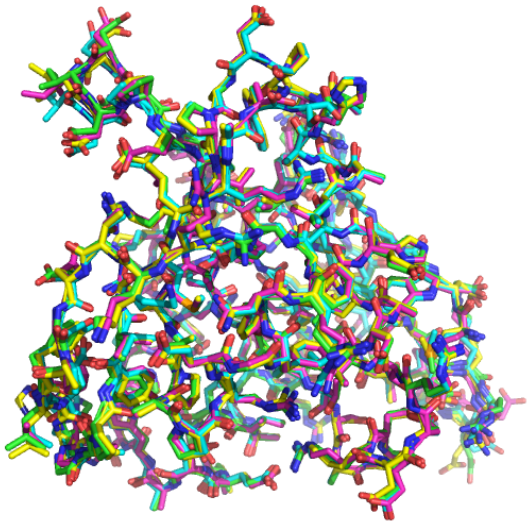
Figure 6.1 Cartoon representation of the crystal structure of BPSL0606 in the asymmetric unit, consisting of 4 protein chains which are denoted by A (green), B (cyan), C (magenta) and D (yellow). This figure was produced using Pymol (Schrodinger, 2010).

Analysis of the 4 subunits was performed by LSQKAB in the CCP4 suite (Kabsch, 1976). An average of RMSD of Cα between each subunit was calculated in the range between 0.33 Å and 0.41 Å with maximum XYZ displacement of 1.25-2.26 Å, summarised in Table 6.1.

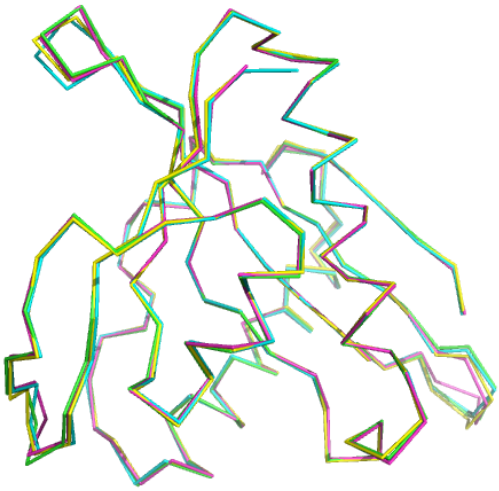
Table 6.1 Analysis of XYZ displacement between 4 subunits in the BPSL0606 structure by LSQKAB

Chain	**XYZ displacement (Å)**	
	RMSD	Maximum
A-B	0.382	2.256
A-C	0.34	1.739
A-D	0.34	1.246
B-C	0.41	1.863
B-D	0.33	1.529
C-D	0.394	2.023

The main chain Cαs of the 4 subunits in the BPSL0606 structure were superimposed (Figure 6.2). For the side chain of these subunits, there was a slightly different orientation of some residues. These included the residues His 140 in chain C and D, which had poor electron density.



a)



b)

Figure 6.2 An overlap of the 4 chains of BPSL0606 seen in the crystal structure: Chains A, B, C and D are colored in green, cyan, magenta and yellow, respectively and shown as sticks (a) and as ribbons (b). This figure was produced using Pymol (Schrodinger, 2010).

6.1.1. Molecular surfaces

To analyse the protein surface, interface and assemblies involving the stability of macromolecules in the crystal form, coordinates of the structure model were input into the PISA server via http://pdbe.org/pisa/ (Krissinel, 2009; Krissinel & Henrick, 2007). The results revealed the structural information of the individual protein chains and also of the molecular assemblies.

The solvent accessible areas of each of the four subunits ranged from 8740 Å^2 to 8950 Å^2 with an average of 8840 Å^2, with approximately 140/151 residues contributing solvent exposed atoms (Table 6.2). As 100 Å^2 is the approximate surface area for a single methionine side chain, for example, the surface areas for each molecule are very similar.

Table 6.2 Analysis of surface areas of each subunit in the BPSL0606 structure

##			Range	Class	Structure		Surface			ΔG, kcal/mol
Id	NN	«»			Nat	Nres	sNat	sNres	Area, Å2	
1	1	⊙	A	Protein	1197	151	727	141	8860.6	-126.7
	2	○	B	Protein	1203	152	729	141	8736.2	-125.2
	3	○	C	Protein	1197	151	737	143	8841.3	-127.2
	4	○	D	Protein	1197	151	735	140	8931.3	-123.5
				Average:	1198	151	732	141	8842.3	-125.7

6.1.2. Dimer interfaces and assemblies

In the asymmetric unit of the BPSL0606 crystal, two dimer interfaces were found, in which chain A formed an interface with chain B, and chain C formed an interface with chain D.

Table 6.3 Analysis of dimer interface in the BPSL0606 structure

##			Structure 1			×	Structure 2					interface	ΔiG	ΔiG	NHB	NSB	NDS	CSS
Id	NN	«»	Range	iNat	iNres		Range	Symmetry op-n	Sym.ID	iNat	iNres	area, Å2	kcal/mol	P-value				
1	1	⊙	D	103	25	◊	C	x,y,z	1_555	102	25	983.3	-13.3	0.118	15	0	0	1.000
	2	○	B	103	25	◊	A	x,y,z	1_555	100	26	955.9	-14.1	0.092	13	0	0	1.000
											Average:	969.6	-13.7	0.105	14	0	0	1.000

Within the asymmetric unit, the four chains come together to form two dimers A-B and C-D (Figure 6.1). A superposition of the C-D dimer onto the A-B dimer shows a mean Cα RMSD of 0.32 Å^2, indicating that these two dimers are very similar (Figure 6.3).

Figure 6.3 Superposition of the A-B and C-D dimers: The dimer A-B is closely related to the dimer C-D. This figure was produced using Pymol (Schrodinger, 2010).

The dimer interface is constructed from one face of the α-helix α_1, and the loop to the strand β_2 packing against the equivalent residues for the 2-fold related subunit (Figure 6.3). As this proposed dimer interface is not particularly extensive, the buried surface area on dimer formation was investigated.

The interface involves 25 residues in chain A, interacting with the same residues in chain B (Table 6.3), with an average interface area of 960 $Å^2$, which is approximately 11% of the subunit surface area. This value is within the range of 6-29% of buried surface area seen in a survey of dimeric proteins (Jones & Thornton, 1995). The dimer formation in the crystal packing agrees with the gel filtration result, which predicted that BPSL0606 is likely to be a dimeric protein, not a monomer.

13 potential hydrogen bonds across the interface were identified, as well as a number of hydrophobic interactions (Table 6.3). The residues involved in this interface in chain A and chain B are shown in Table 6.4 and the interactions made by these residues are described in Table 6.5.

A similar analysis was undertaken for the C-D dimer, which is very similar to the A-B dimer and shown in Table 6.6 and 6.7.

Table 6.4 Analysis of the A-B dimer interfacing residues of the BPSL0606 structure: Interfacing residues in Chain A (right) and Chain B (left) are highlighted in yellow.

Interfacing residues (not a contact table) XML Display level: Residues

Inaccessible residues
Solvent-accessible residues
HSDC Residues making Hydrogen/Disulphide bond, Salt bridge or Covalent link
Interfacing residues
ASA Accessible Surface Area, Å² BSA Buried Surface Area, Å² ΔiG Solvation energy effect, kcal/mol |||| Buried area percentage, one bar per 10%

##	Structure 1	HSDC	ASA	BSA	ΔiG
1	B:MSE 1		98.91	0.00	0.00
2	B:ASP 2		71.11	0.00	0.00
3	B:TRP 3		57.55	0.00	0.00
4	B:THR 4		44.12	0.00	0.00
5	B:CYS 5		48.25	0.00	0.00
6	B:CYS 6		16.36	0.00	0.00
7	B:GLU 7		51.33	0.00	0.00
8	B:PHE 8		61.75	52.67 ||||||||	0.76
9	B:ARG 9		178.31	5.59 |	0.08
10	B:HIS 10		114.91	0.00	0.00
11	B:LEU 11		20.17	10.99 |||||	-0.09
12	B:SER 12		54.24	0.47 |	0.01
13	B:SER 13		99.50	70.15 |||||||	0.37
14	B:ASN 14		78.76	14.54 ||	-0.05
15	B:GLU 15		42.01	0.00	0.00
16	B:LEU 16		48.48	45.96 |||||||||	0.74
17	B:TYR 17	H	143.02	130.20 |||||||||	0.82
18	B:MSE 18		24.72	0.00	0.00
19	B:ILE 19		1.66	0.00	0.00
20	B:LEU 20		40.49	39.34 |||||||||	0.59
21	B:ARG 21	H	162.93	77.90 |||||	-0.73
22	B:THR 22		18.20	0.00	0.00
23	B:ARG 23		0.74	0.00	0.00
24	B:ASN 24	H	72.04	60.25 ||||||||	-0.56
25	B:ALA 25		49.90	9.95 ||	0.14
26	B:VAL 26		16.24	0.00	0.00
27	B:LEU 27		38.11	0.00	0.00
28	B:VAL 28		31.77	26.92 ||||||||	0.42
29	B:VAL 29		109.19	45.32 |||||	0.70
30	B:GLU 30		90.54	0.00	0.00
31	B:ASP 31		97.14	0.00	0.00
32	B:ALA 32		98.90	11.12 ||	0.18
33	B:HIS 33		50.77	2.64 |	-0.02
34	B:THR 34	H	138.11	90.49 |||||||	0.58
35	B:HIS 35	H	79.25	24.71 ||||	-0.04
36	B:LEU 36		134.27	66.54 |||||	1.06
37	B:ASP 37		7.87	4.29 |||||	-0.00
38	B:ILE 38	H	101.91	97.05 |||||||||	1.34
39	B:ASP 39	H	46.18	8.98 ||	-0.02
40	B:GLY 40	H	66.40	31.50 |||||	0.45
41	B:LYS 41		108.91	0.00	0.00
42	B:ASP 42	H	12.73	11.78 |||||||||	0.33
43	B:GLU 43		106.61	14.09 ||	0.02
44	B:PHE 44		164.94	0.00	0.00

##	Structure 2	HSDC	ASA	BSA	ΔiG
1	A:MSE 1		92.18	0.00	0.00
2	A:ASP 2		69.91	0.00	0.00
3	A:TRP 3		65.87	0.00	0.00
4	A:THR 4		43.65	0.00	0.00
5	A:CYS 5		48.48	0.00	0.00
6	A:CYS 6		14.67	0.00	0.00
7	A:GLU 7		65.01	0.00	0.00
8	A:PHE 8		64.68	50.65 ||||||||	0.74
9	A:ARG 9		178.23	6.25 |	0.07
10	A:HIS 10		107.98	0.00	0.00
11	A:LEU 11		21.02	12.12 ||||||	-0.12
12	A:SER 12		51.75	0.15 |	0.00
13	A:SER 13		90.79	65.24 |||||||	0.24
14	A:ASN 14		106.42	30.28 |||	-0.32
15	A:GLU 15		51.06	0.00	0.00
16	A:LEU 16		52.87	49.21 |||||||||	0.79
17	A:TYR 17	H	148.33	128.69 ||||||||	0.87
18	A:MSE 18		34.08	0.00	0.00
19	A:ILE 19		1.92	0.00	0.00
20	A:LEU 20		40.15	39.20 |||||||||	0.60
21	A:ARG 21	H	165.39	77.00 |||||	-0.62
22	A:THR 22		19.01	0.00	0.00
23	A:ARG 23		0.44	0.00	0.00
24	A:ASN 24	H	74.53	62.80 ||||||||	-0.58
25	A:ALA 25		48.71	9.85 |||	0.15
26	A:VAL 26		15.55	0.00	0.00
27	A:LEU 27		31.19	0.00	0.00
28	A:VAL 28		29.62	25.26 |||||||||	0.40
29	A:VAL 29		99.89	41.62 |||||	0.64
30	A:GLU 30		87.16	0.00	0.00
31	A:ASP 31		87.24	0.00	0.00
32	A:ALA 32		99.22	10.71 ||	0.16
33	A:HIS 33		44.63	2.62 |	-0.03
34	A:THR 34	H	139.40	86.40 |||||||	0.70
35	A:HIS 35	H	68.72	22.79 ||||	-0.04
36	A:LEU 36		141.77	77.68 ||||||	1.16
37	A:ASP 37		6.21	3.92 ||||||	0.01
38	A:ILE 38	H	97.71	92.68 |||||||||	1.33
39	A:ASP 39	H	43.23	8.57 ||	-0.01
40	A:GLY 40	H	71.02	35.68 ||||||	0.48
41	A:LYS 41		114.80	0.00	0.00
42	A:ASP 42	H	13.67	11.73 |||||||||	0.34
43	A:GLU 43		99.85	6.87 |	0.05
44	A:PHE 44		164.53	0.00	0.00

Table 6.5 Interaction of H-bond between Chain A and Chain B of the BPSL0606 structure: 13 H-bonds were found in the A-B interface.

Hydrogen bonds XML

##	Structure 1	Dist. [Å]	Structure 2
1	B:TYR 17[OH]	2.52	A:ASP 42[OD2]
2	B:ARG 21[NE]	3.44	A:ILE 38[O]
3	B:ASN 24[ND2]	3.70	A:HIS 35[O]
4	B:THR 34[OG1]	2.50	A:THR 34[OG1]
5	B:ASP 39[N]	3.58	A:TYR 17[OH]
6	B:GLY 40[N]	2.94	A:TYR 17[OH]
7	B:TYR 17[OH]	3.55	A:ASP 39[N]
8	B:TYR 17[OH]	3.83	A:ASP 42[N]
9	B:TYR 17[OH]	2.92	A:GLY 40[N]
10	B:THR 34[OG1]	3.39	A:THR 34[N]
11	B:HIS 35[O]	3.60	A:ASN 24[ND2]
12	B:ILE 38[O]	3.33	A:ARG 21[NE]
13	B:ASP 42[OD2]	2.65	A:TYR 17[OH]

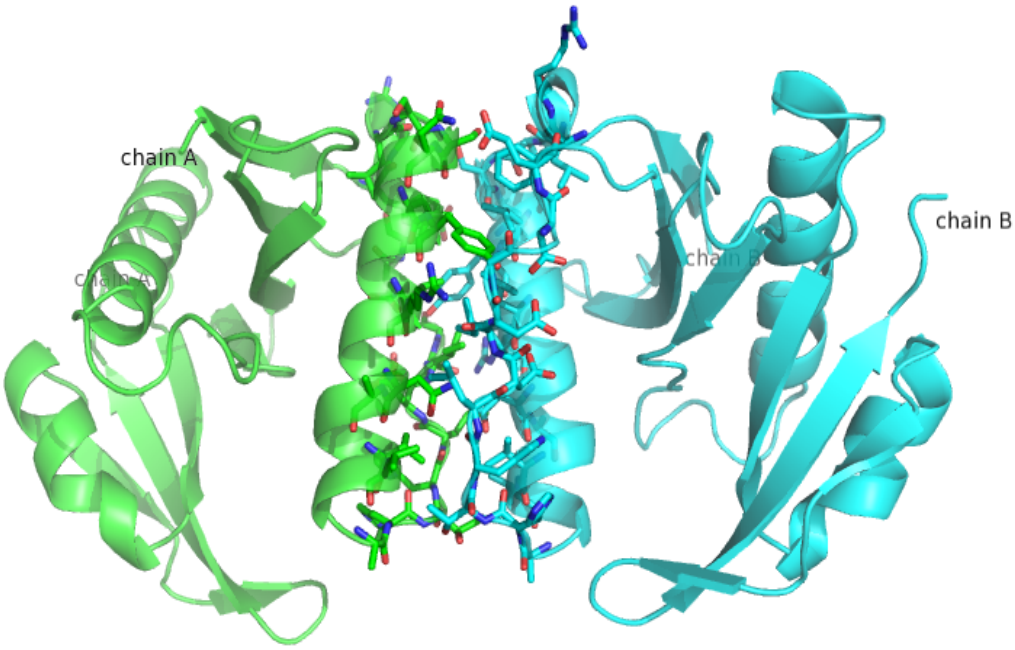


Figure 6.4 Residues forming the interface of the A-B dimer: The dimer interface is formed by 25 residues on the α-helix α1 and the loop connected the strand β2 in one chain against the equivalent residues in the other. This figure was produced using Pymol (Schrodinger, 2010).

Table 6.6 Analysis of the C-D dimer interfacing residues of the BPSL0606 structure: Interfacing residues in Chain C (right) and Chain D (left) are highlighted in yellow.

Interfacing residues (not a contact table) XML Display level: Residues

Inaccessible residues
Solvent-accessible residues
HSDC Residues making Hydrogen/Disulphide bond, Salt bridge or Covalent link
Interfacing residues
ASA Accessible Surface Area, Å² **BSA** Buried Surface Area, Å² **ΔiG** Solvation energy effect, kcal/mol |||| Buried area percentage, one bar per 10%

##	Structure 1	HSDC	ASA	BSA	ΔiG
1	D:MSE 1		94.54	0.00	0.00
2	D:ASP 2		90.59	0.00	0.00
3	D:TRP 3		61.33	0.00	0.00
4	D:THR 4		46.70	0.00	0.00
5	D:CYS 5		48.56	0.00	0.00
6	D:CYS 6		16.13	0.00	0.00
7	D:GLU 7		51.43	0.00	0.00
8	D:PHE 8		61.66	54.14 |||||||||	0.81
9	D:ARG 9		179.37	5.82 |	0.07
10	D:HIS 10		110.03	0.00	0.00
11	D:LEU 11		20.78	10.45 ||||||	-0.09
12	D:SER 12		53.61	0.94 |	0.02
13	D:SER 13		88.63	64.29 ||||||||	0.23
14	D:ASN 14		105.83	31.89 ||||	-0.33
15	D:GLU 15		54.52	0.00	0.00
16	D:LEU 16		49.46	46.12 ||||||||||	0.74
17	D:TYR 17	H	148.50	129.07 |||||||||	0.84
18	D:MSE 18		41.53	0.00	0.00
19	D:ILE 19		1.59	0.00	0.00
20	D:LEU 20		36.62	35.51 ||||||||||	0.54
21	D:ARG 21		161.97	74.18 |||||	-0.67
22	D:THR 22		18.05	0.00	0.00
23	D:ARG 23		0.73	0.00	0.00
24	D:ASN 24		76.45	65.23 |||||||||	-0.54
25	D:ALA 25		50.95	12.34 |||	0.19
26	D:VAL 26		15.24	0.00	0.00
27	D:LEU 27		34.30	0.00	0.00
28	D:VAL 28		32.32	28.46 |||||||||	0.43
29	D:VAL 29		106.38	48.04 |||||	0.72
30	D:GLU 30		81.23	0.00	0.00
31	D:ASP 31		90.65	0.00	0.00
32	D:ALA 32		93.51	11.79 ||	0.19
33	D:HIS 33		85.62	1.04 |	-0.00
34	D:THR 34	H	136.60	87.80 ||||||||	0.62
35	D:HIS 35	H	69.93	27.65 ||||	0.01
36	D:LEU 36		135.42	75.12 ||||||	1.11
37	D:ASP 37	H	7.77	4.23 ||||||	0.02
38	D:ILE 38	H	103.92	97.89 ||||||||||	1.36
39	D:ASP 39	H	45.19	7.12 ||	-0.03
40	D:GLY 40	H	71.10	33.50 |||||	0.48
41	D:LYS 41		117.43	0.00	0.00
42	D:ASP 42	H	11.67	10.71 ||||||||||	0.32
43	D:GLU 43		98.89	21.91 |||	-0.12
44	D:PHE 44		168.45	0.00	0.00

##	Structure 2	HSDC	ASA	BSA	ΔiG
1	C:MSE 1		98.66	0.00	0.00
2	C:ASP 2		73.37	0.00	0.00
3	C:TRP 3		65.21	0.00	0.00
4	C:THR 4		46.32	0.00	0.00
5	C:CYS 5		47.73	0.00	0.00
6	C:CYS 6		13.14	0.00	0.00
7	C:GLU 7		74.51	0.00	0.00
8	C:PHE 8		66.86	56.61 |||||||||	0.81
9	C:ARG 9		182.93	3.86 |	0.02
10	C:HIS 10		98.81	0.00	0.00
11	C:LEU 11		22.07	11.73 ||||||	-0.09
12	C:SER 12		52.93	0.90 |	0.01
13	C:SER 13		96.92	65.11 |||||||	0.26
14	C:ASN 14		113.40	37.17 ||||	-0.51
15	C:GLU 15		56.07	0.00	0.00
16	C:LEU 16		51.71	48.20 ||||||||||	0.77
17	C:TYR 17	H	146.21	129.16 |||||||||	0.81
18	C:MSE 18		36.84	0.00	0.00
19	C:ILE 19		1.71	0.00	0.00
20	C:LEU 20		35.53	34.46 ||||||||||	0.52
21	C:ARG 21	H	162.69	74.04 |||||	-0.74
22	C:THR 22		19.92	0.00	0.00
23	C:ARG 23		0.58	0.00	0.00
24	C:ASN 24	H	71.65	60.95 ||||||||	-0.56
25	C:ALA 25		51.05	10.98 |||	0.16
26	C:VAL 26		14.06	0.00	0.00
27	C:LEU 27		34.02	0.00	0.00
28	C:VAL 28		32.69	29.50 ||||||||||	0.45
29	C:VAL 29		102.38	43.96 |||||	0.67
30	C:GLU 30		88.45	0.00	0.00
31	C:ASP 31		88.13	0.00	0.00
32	C:ALA 32		84.57	10.67 ||	0.17
33	C:HIS 33		48.29	1.60 |	-0.02
34	C:THR 34	H	132.35	86.03 |||||||	0.42
35	C:HIS 35		72.41	26.35 ||||	-0.03
36	C:LEU 36		139.57	81.45 ||||||	1.24
37	C:ASP 37		6.87	3.80 ||||||	0.02
38	C:ILE 38		101.45	96.76 ||||||||||	1.37
39	C:ASP 39	H	39.00	6.76 ||	-0.01
40	C:GLY 40	H	68.68	34.66 ||||||	0.46
41	C:LYS 41		106.66	0.00	0.00
42	C:ASP 42	H	12.52	11.24 ||||||||	0.36
43	C:GLU 43		103.37	15.44 ||	-0.19
44	C:PHE 44		163.11	0.00	0.00

Table 6.7 Interaction of H-bond between Chain C and Chain D of the BPSL0606 structure: 15 H-bonds were found in the C-D interface.

Hydrogen bonds XML

##	Structure 1	Dist. [Å]	Structure 2
1	D:ASP 39[N]	3.64	C:TYR 17[OH]
2	D:GLY 40[N]	3.07	C:TYR 17[OH]
3	D:ASP 42[N]	3.89	C:TYR 17[OH]
4	D:ILE 38[N]	3.40	C:ASN 24[OD1]
5	D:ASP 37[N]	3.17	C:ASN 24[OD1]
6	D:THR 34[N]	3.35	C:THR 34[OG1]
7	D:THR 34[OG1]	2.61	C:THR 34[OG1]
8	D:TYR 17[OH]	2.61	C:ASP 42[OD2]
9	D:TYR 17[OH]	3.81	C:ASP 42[N]
10	D:TYR 17[OH]	2.88	C:GLY 40[N]
11	D:TYR 17[OH]	3.49	C:ASP 39[N]
12	D:THR 34[OG1]	3.38	C:THR 34[N]
13	D:HIS 35[O]	3.19	C:ASN 24[ND2]
14	D:ILE 38[O]	3.33	C:ARG 21[NE]
15	D:ASP 42[OD2]	2.37	C:TYR 17[OH]

6.2. The crystal structure of the truncated BPSL0606 contained 2 molecules in the asymmetric unit ($P2_12_12_1$ crystal form)

In the second, $P2_12_12_1$ crystal form of BPSL0606, two subunits are present in the asymmetric unit, which are also arranged in the same orientation as the dimer seen in crystal form one (Figure 6.5).

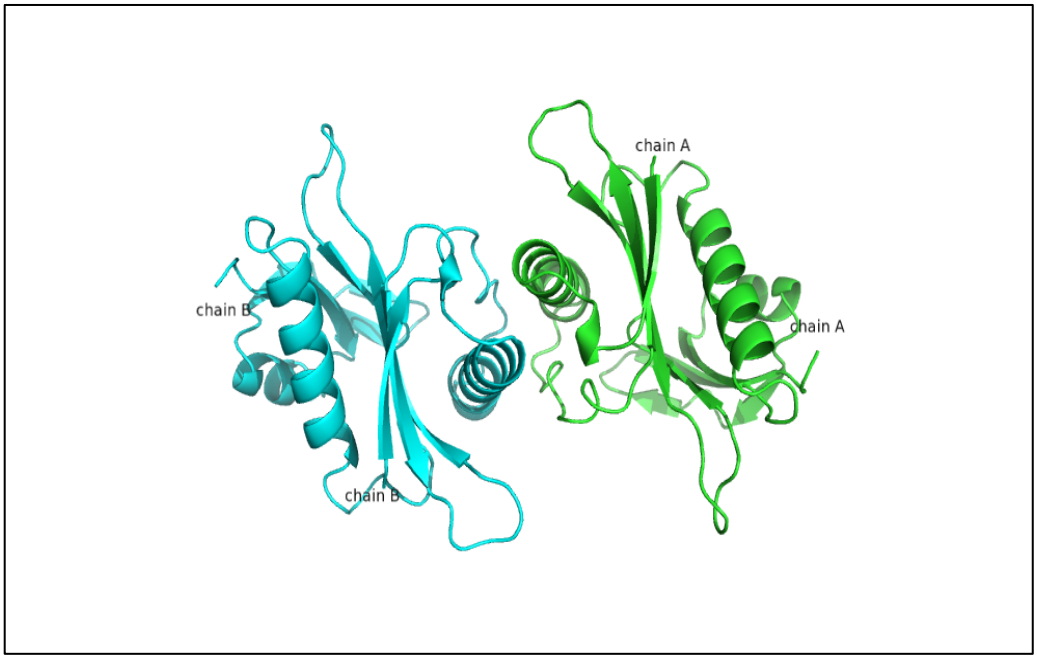


Figure 6.5 Cartoon representation of the crystal structure of the truncated BPSL0606 in the asymmetric unit consisting of 2 protein chains, which are designated by A (green) and B (cyan). This figure was produced using Pymol (Schrodinger, 2010).

A superposition of each subunit shows the average RMSD of Cα was 0.55 Å with a maximum XYZ displacement of 2.56 Å, calculated using LSQKAB in the CCP4 suite (Kabsch, 1976). The main chains Cα of these two subunits in the truncated BPSL0606 structure were similarly oriented (Figure 6.6). A slight difference in side chain orientation was observed. Since residue Glu 56 in Chain A has poor electron density, its side chain was displaced from the equivalent residue in Chain B, resulting in it being in an outlier region of the Ramachandran plot.

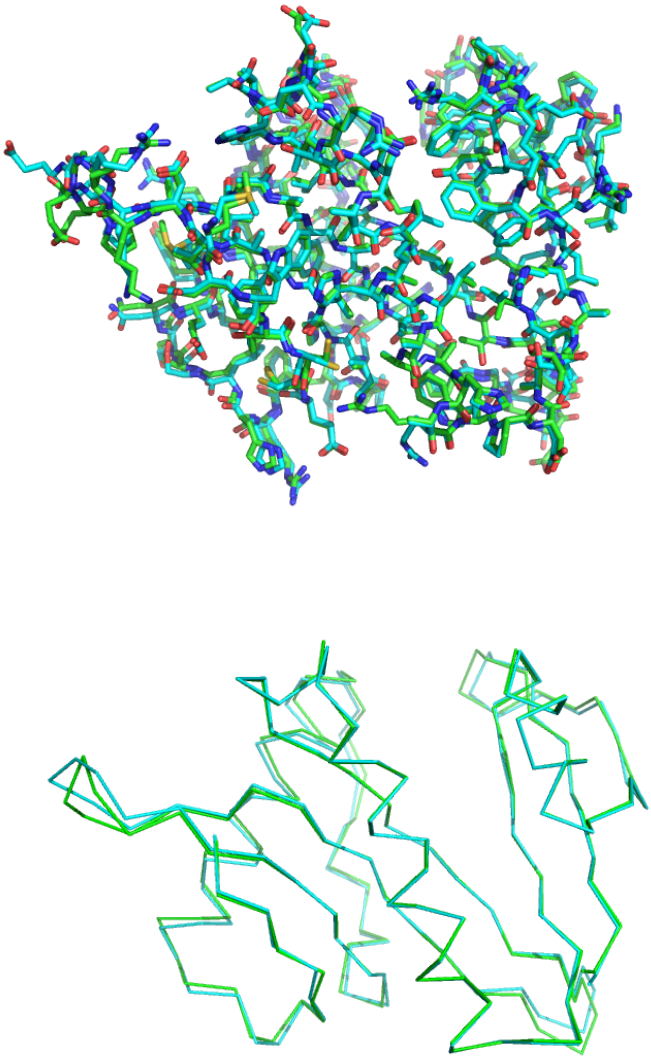
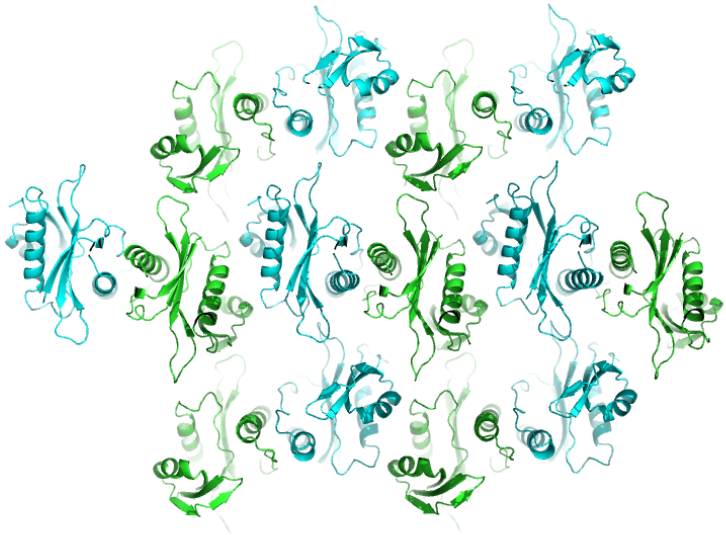
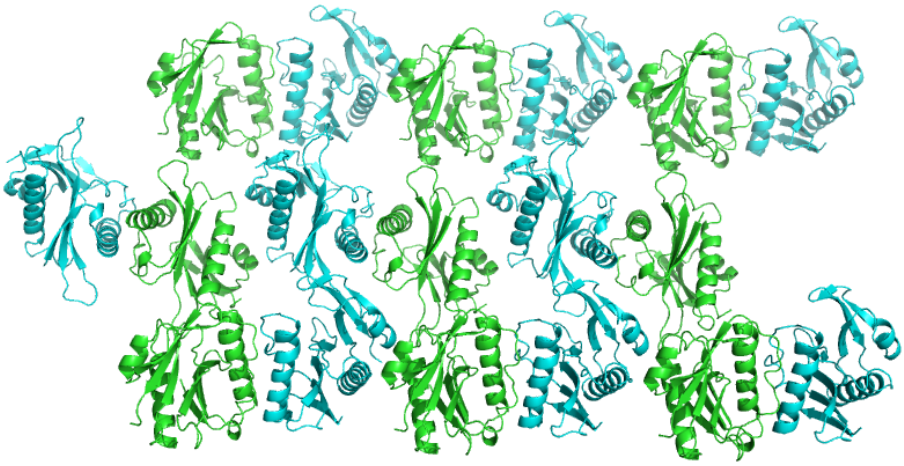


Figure 6.6 An overlap of 2 subunits in the truncated BPSL0606 structure: Chain A is colored in green and chain B is colored in cyan. This figure was produced using Pymol (Schrodinger, 2010).

The crystal packing between the dimers in the $P2_12_12_1$ and $P2_1$ forms is similar within one layer but quite different in the other layer as can be seen in Figure 6.7.



a)



b)

Figure 6.7 Crystal packing of BPSL0606 crystals in $P2_1$ and $P2_12_12_1$: Symmetry related molecules of the dimer (a) in the $P2_1$ crystal form containing 4 molecules in the asymmetric unit and (b) in the $P2_12_12_1$ crystal form containing 2 molecules in the asymmetric unit. This figure was produced using Pymol (Schrodinger, 2010).

The molecular 2-fold interface in the $P2_12_12_1$ crystal form is built for the same residues on the α-helix α1 and the loop connected the strand β2 as in the $P2_1$ crystal form with similar interactions being made (Table 6.8, 6.9 and 6.10).

Table 6.8 Analysis of surface areas of each subunit in the truncated BPSL0606 structure

##			Range	Class	Structure		Surface			ΔG, kcal/mol
Id	NN	«»			Nat	Nres	sNat	sNres	Area, Å2	
1	1	⊙	A	Protein	1203	152	723	142	8655.9	−129.4
	2	○	B	Protein	1203	152	745	142	8875.9	−129.2
				Average:	1203	152	734	142	8765.9	−129.3

Table 6.9 Analysis of A-B dimer interfacing residues of the truncated BPSL0606 structure: Interfacing residues in Chain A (right) and Chain B (left) are highlighted in yellow.

Interfacing residues (not a contact table) XML Display level: Residues

Inaccessible residues
Solvent-accessible residues
HSDC Residues making Hydrogen/Disulphide bond, Salt bridge or Covalent link
Interfacing residues
ASA Accessible Surface Area, Å² **BSA** Buried Surface Area, Å² **ΔiG** Solvation energy effect, kcal/mol |||| Buried area percentage, one bar per 10%

##	Structure 1	HSDC	ASA	BSA	ΔiG
1	B:MET 1		95.41	0.00	0.00
2	B:ASP 2		93.85	0.00	0.00
3	B:TRP 3		66.17	0.00	0.00
4	B:THR 4		34.93	0.00	0.00
5	B:CYS 5		70.69	0.00	0.00
6	B:CYS 6		23.61	0.00	0.00
7	B:GLU 7		58.69	0.00	0.00
8	B:PHE 8		66.80	54.94 |||||||||	0.78
9	B:ARG 9		185.63	1.67 |	0.03
10	B:HIS 10		105.47	0.00	0.00
11	B:LEU 11		21.13	9.48 |||||	−0.07
12	B:SER 12		55.85	1.53 |	0.02
13	B:SER 13		89.85	64.61 ||||||||	0.17
14	B:ASN 14		123.60	35.56 |||	−0.27
15	B:GLU 15		49.41	0.00	0.00
16	B:LEU 16		46.19	45.36 ||||||||||	0.73
17	B:TYR 17	H	153.78	146.74 ||||||||||	1.14
18	B:MET 18		33.19	0.00	0.00
19	B:ILE 19		0.87	0.00	0.00
20	B:LEU 20		36.53	35.43 ||||||||||	0.53
21	B:ARG 21		161.65	75.53 |||||	−0.89
22	B:THR 22		17.70	0.00	0.00
23	B:ARG 23		0.61	0.00	0.00
24	B:ASN 24		77.42	64.72 |||||||||	−0.57
25	B:ALA 25		50.71	12.51 |||	0.19
26	B:VAL 26		18.19	0.00	0.00
27	B:LEU 27		31.23	0.00	0.00
28	B:VAL 28		27.94	22.41 |||||||||	0.35
29	B:VAL 29		110.00	49.22 |||||	0.74
30	B:GLU 30		89.62	0.00	0.00
31	B:ASP 31		84.92	0.00	0.00
32	B:ALA 32		99.97	9.04 |	0.14
33	B:HIS 33		60.86	2.08 |	−0.01
34	B:THR 34	H	135.65	88.03 ||||||	0.36
35	B:HIS 35	H	77.10	26.64 ||||	−0.01
36	B:LEU 36		141.11	77.12 ||||||	1.19
37	B:ASP 37		6.59	3.42 ||||||	0.01
38	B:ILE 38	H	101.44	95.59 ||||||||||	1.35
39	B:ASP 39	H	48.31	10.68 |||	−0.03
40	B:GLY 40	H	55.76	32.69 ||||||	0.45
41	B:LYS 41		106.83	0.00	0.00
42	B:ASP 42	H	10.60	10.29 ||||||||||	0.31
43	B:GLU 43		112.90	21.00 ||	0.02

##	Structure 2	HSDC	ASA	BSA	ΔiG
1	A:MET 1		99.00	0.00	0.00
2	A:ASP 2		71.22	0.00	0.00
3	A:TRP 3		66.62	0.00	0.00
4	A:THR 4		43.84	0.00	0.00
5	A:CYS 5		43.65	0.00	0.00
6	A:CYS 6		15.19	0.00	0.00
7	A:GLU 7		72.15	0.00	0.00
8	A:PHE 8		47.88	40.22 |||||||||	0.59
9	A:ARG 9		166.50	1.67 |	0.03
10	A:HIS 10		111.21	0.00	0.00
11	A:LEU 11		22.62	12.71 ||||||	−0.10
12	A:SER 12		53.21	0.78 |	0.01
13	A:SER 13		95.37	67.75 ||||||||	0.17
14	A:ASN 14		118.61	29.34 |||	−0.33
15	A:GLU 15		56.06	0.00	0.00
16	A:LEU 16		47.21	45.88 ||||||||||	0.73
17	A:TYR 17	H	149.69	129.02 ||||||||||	0.84
18	A:MET 18		46.38	0.00	0.00
19	A:ILE 19		1.29	0.00	0.00
20	A:LEU 20		39.87	38.64 ||||||||||	0.58
21	A:ARG 21	H	161.26	77.25 |||||	−0.90
22	A:THR 22		19.63	0.00	0.00
23	A:ARG 23		0.46	0.00	0.00
24	A:ASN 24	H	76.13	65.06 |||||||||	−0.59
25	A:ALA 25		52.46	12.34 |||	0.19
26	A:VAL 26		16.87	0.00	0.00
27	A:LEU 27		33.36	0.00	0.00
28	A:VAL 28		29.51	24.50 |||||||||	0.39
29	A:VAL 29		102.96	44.18 |||||	0.69
30	A:GLU 30		90.00	0.00	0.00
31	A:ASP 31		87.46	0.00	0.00
32	A:ALA 32		94.39	10.85 ||	0.17
33	A:HIS 33		40.57	1.16 |	−0.01
34	A:THR 34	H	134.53	83.48 |||||||	0.34
35	A:HIS 35		73.94	25.72 ||||	−0.03
36	A:LEU 36		132.94	74.34 ||||||	1.16
37	A:ASP 37		7.86	4.24 ||||||	0.02
38	A:ILE 38		101.78	97.16 ||||||||||	1.30
39	A:ASP 39	H	45.29	10.18 |||	−0.01
40	A:GLY 40	H	46.46	28.60 |||||||	0.42
41	A:LYS 41		106.44	0.00	0.00
42	A:ASP 42	H	9.11	8.65 ||||||||||	0.35
43	A:GLU 43		98.22	49.70 ||||||	−0.46

Table 6.10 Interaction of H-bond between Chain A and Chain B of the truncated BPSL0606 structure: 11 H-bonds made by interacting residues in Chain A (right) and Chain B (left) found.

Hydrogen bonds XML

##	Structure 1	Dist. [Å]	Structure 2
1	B:ASP 42[N]	3.87	A:TYR 17[OH]
2	B:ASP 39[N]	3.43	A:TYR 17[OH]
3	B:GLY 40[N]	2.83	A:TYR 17[OH]
4	B:THR 34[OG1]	2.59	A:THR 34[OG1]
5	B:TYR 17[OH]	2.60	A:ASP 42[OD2]
6	B:TYR 17[OH]	3.63	A:ASP 42[N]
7	B:TYR 17[OH]	3.60	A:ASP 39[N]
8	B:TYR 17[OH]	2.86	A:GLY 40[N]
9	B:HIS 35[O]	3.53	A:ASN 24[ND2]
10	B:ILE 38[O]	3.33	A:ARG 21[NE]
11	B:ASP 42[OD2]	2.51	A:TYR 17[OH]

6.3. Interface similarity

To investigate whether the observed BPSL0606 dimer is biologically relevant, the dimer structure was searched against known structures to find any similar oligomeric assemblies, using the PISA server (Krissinel, 2009; Krissinel & Henrick, 2007). The results are shown in Table 6.11 and 6.12.

Table 6.11 Interface similarity of the A-B dimer (a) and the C-D dimer (b) in the crystal form $P2_1$.

a) A-B interface similarity

##	Entry	Intf/mm No	Size	Space group	Q score	Seq. Id	Interface area, Å2	ΔiG kcal/mol	CSS	Title
1	1xeb	1	2	P 32 2 1	0.827	0.366	956.5	-19.2	1.000	CRYSTAL STRUCTURE OF AN ACYL-COA N-ACYLTRANSFERASE FROM PSEUDOMONAS AERUGINOSA
2	3efa	1	2	P 65 2 2	0.646	0.228	840.3	-7.5	0.630	CRYSTAL STRUCTURE OF PUTATIVE N-ACETYLTRANSFERASE FROM LACTOBACILLUS PLANTARUM
3	1q2y	1	1	P 21 21 21	0.613	0.164	376.0	-0.5	0.000	CRYSTAL STRUCTURE OF THE PROTEIN YJCF FROM BACILLUS SUBTILIS: A MEMBER OF THE GCN5-RELATED N-ACETYLTRANSFERASE SUPERFAMILY FOLD
4	3p2h	1	1	P 32 2 1	0.489	0.162	521.0	-3.3	-0.000	CRYSTAL STRUCTURE OF TOFI IN A TERNARY COMPLEX WITH AN INHIBITOR AND MTA
5	2jdd	1	1	C 1 2 1	0.489	0.142	758.8	-4.1	-0.000	GLYPHOSATE N-ACETYLTRANSFERASE BOUND TO ACETYL COA AND 3-PHOSPHOGLYCERATE
6	2cnm	3	6	H 3 2	0.487	0.156	543.6	-3.2	1.000	RIMI - RIBOSOMAL S18 N-ALPHA-PROTEIN ACETYLTRANSFERASE IN COMPLEX WITH A BISUBSTRATE INHIBITOR (CTERM-ARG-ARG-PHE- TYR-ARG-ALA-N-ALPHA-ACETYL-S-COA).
7	2cns	6	1	H 3 2	0.486	0.154	496.7	-4.0	0.033	RIMI - RIBOSOMAL S18 N-ALPHA-PROTEIN ACETYLTRANSFERASE IN COMPLEX WITH ACETYLCOA.
8	3p2f	1	1	P 32 2 1	0.484	0.167	498.3	-3.1	0.000	CRYSTAL STRUCTURE OF TOFI IN AN APO FORM
9	2jdc	1	1	C 1 2 1	0.478	0.143	857.8	-5.4	0.014	GLYPHOSATE N-ACETYLTRANSFERASE BOUND TO OXIDIZED COA AND SULFATE
10	2bsw	1	2	C 1 2 1	0.477	0.153	849.1	-4.7	0.134	CRYSTAL STRUCTURE OF A GLYPHOSATE-N-ACETYLTRANSFERASE OBTAINED BY DNA SHUFFLING.
11	1kux	2	1	C 2 2 21	0.470	0.104	543.3	-4.6	-0.000	X-RAY CRYSTALLOGRAPHIC STUDIES OF SEROTONIN N- ACETYLTRANSFERASE CATALYSIS AND INHIBITION
12	1kuv	2	1	C 2 2 21	0.464	0.118	534.7	-7.1	-0.000	X-RAY CRYSTALLOGRAPHIC STUDIES OF SEROTONIN N- ACETYLTRANSFERASE CATALYSIS AND INHIBITION
13	1kuy	2	1	C 2 2 21	0.460	0.109	521.2	-6.4	-0.000	X-RAY CRYSTALLOGRAPHIC STUDIES OF SEROTONIN N- ACETYLTRANSFERASE CATALYSIS AND INHIBITION
14	2i79	1	2	P 21 21 2	0.458	0.116	1482.1	-2.2	0.462	THE CRYSTAL STRUCTURE OF THE ACETYLTRANSFERASE OF GNAT FAMILY FROM STREPTOCOCCUS PNEUMONIAE
15	2psw	4	1	P 1 21 1	0.458	0.107	543.2	-3.1	-0.000	HUMAN MAK3 HOMOLOG IN COMPLEX WITH COA
16	3tfy	4	2	P 1 21 1	0.453	0.115	545.6	-0.7	-0.000	NAA50P AMINO-TERMINAL ACETYLTRANSFERASE BOUND TO SUBSTRATE PEPTIDE FRAGMENT AND COA
17	1cjw	2	1	C 2 2 21	0.451	0.103	549.7	-5.3	0.014	SEROTONIN N-ACETYLTRANSFERASE COMPLEXED WITH A BISUBSTRATE ANALOG
18	3d8p	1	1	I 4 2 2	0.450	0.125	418.7	3.9	-0.000	CRYSTAL STRUCTURE OF ACETYLTRANSFERASE OF GNAT FAMILY (NP_373092.1) FROM STAPHYLOCOCCUS AUREUS MU50 AT 2.20 A RESOLUTION
19	2ge3	2	2	P 21 21 21	0.450	0.151	1034.0	-5.7	0.676	CRYSTAL STRUCTURE OF PROBABLE ACETYLTRANSFERASE FROM AGROBACTERIUM TUMEFACIENS
20	3f8k	1	1	P 21 21 21	0.448	0.110	618.4	-7.2	-0.000	CRYSTAL STRUCTURE OF PROTEIN ACETYLTRANSFERASE (PAT) FROM SULFOLOBUS SOLFATARICUS

b) C-D interface similarity

##	Entry	Intf No	mm Size	Space group	Q score	Seq. Id	Interface area, Å2	ΔG kcal/mol	CSS	Title
1	1xeb	1	2	P 32 2 1	0.829	0.369	956.5	-19.2	1.000	CRYSTAL STRUCTURE OF AN ACYL-COA N-ACYLTRANSFERASE FROM PSEUDOMONAS AERUGINOSA
2	3efa	1	2	P 65 2 2	0.641	0.228	840.3	-7.5	0.630	CRYSTAL STRUCTURE OF PUTATIVE N-ACETYLTRANSFERASE FROM LACTOBACILLUS PLANTARUM
3	1q2y	1	1	P 21 21 21	0.615	0.164	376.0	-0.5	0.000	CRYSTAL STRUCTURE OF THE PROTEIN YJCF FROM BACILLUS SUBTILIS: A MEMBER OF THE GCN5-RELATED N-ACETYLTRANSFERASE SUPERFAMILY FOLD
4	2jdd	1	1	C 1 2 1	0.484	0.143	758.8	-4.1	-0.000	GLYPHOSATE N-ACETYLTRANSFERASE BOUND TO ACETYL COA AND 3-PHOSPHOGLYCERATE
5	2cnm	3	6	H 3 2	0.479	0.150	543.6	-3.2	1.000	RIMI - RIBOSOMAL S18 N-ALPHA-PROTEIN ACETYLTRANSFERASE IN COMPLEX WITH A BISUBSTRATE INHIBITOR (CTERM-ARG-ARG-PHE- TYR-ARG-ALA-N-ALPHA-ACETYL-S-COA).
6	2cns	6	1	H 3 2	0.477	0.150	496.7	-4.0	0.033	RIMI - RIBOSOMAL S18 N-ALPHA-PROTEIN ACETYLTRANSFERASE IN COMPLEX WITH ACETYLCOA.
7	2bsw	1	2	C 1 2 1	0.476	0.151	849.1	-4.7	0.134	CRYSTAL STRUCTURE OF A GLYPHOSATE-N-ACETYLTRANSFERASE OBTAINED BY DNA SHUFFLING.
8	2jdc	1	1	C 1 2 1	0.474	0.144	857.8	-5.4	0.014	GLYPHOSATE N-ACETYLTRANSFERASE BOUND TO OXIDIZED COA AND SULFATE
9	1cjw	2	1	C 2 2 21	0.470	0.110	549.7	-5.3	0.014	SEROTONIN N-ACETYLTRANSFERASE COMPLEXED WITH A BISUBSTRATE ANALOG
10	1kux	2	1	C 2 2 21	0.469	0.116	543.3	-4.6	-0.000	X-RAY CRYSTALLOGRAPHIC STUDIES OF SEROTONIN N- ACETYLTRANSFERASE CATALYSIS AND INHIBITION
11	1kuy	2	1	C 2 2 21	0.463	0.113	521.2	-6.4	-0.000	X-RAY CRYSTALLOGRAPHIC STUDIES OF SEROTONIN N- ACETYLTRANSFERASE CATALYSIS AND INHIBITION
12	2ge3	2	2	P 21 21 21	0.461	0.156	1034.0	-5.7	0.676	CRYSTAL STRUCTURE OF PROBABLE ACETYLTRANSFERASE FROM AGROBACTERIUM TUMEFACIENS
13	2i79	1	2	P 21 21 2	0.453	0.115	1482.1	-2.2	0.462	THE CRYSTAL STRUCTURE OF THE ACETYLTRANSFERASE OF GNAT FAMILY FROM STREPTOCOCCUS PNEUMONIAE
14	1kuv	2	1	C 2 2 21	0.453	0.115	534.7	-7.1	-0.000	X-RAY CRYSTALLOGRAPHIC STUDIES OF SEROTONIN N- ACETYLTRANSFERASE CATALYSIS AND INHIBITION
15	1l0c	2	1	C 2 2 21	0.453	0.101	552.0	-4.2	-0.000	INVESTIGATION OF THE ROLES OF CATALYTIC RESIDUES IN SEROTONIN N-ACETYLTRANSFERASE
16	3f8k	1	1	P 21 21 21	0.452	0.109	618.4	-7.2	-0.000	CRYSTAL STRUCTURE OF PROTEIN ACETYLTRANSFERASE (PAT) FROM SULFOLOBUS SOLFATARICUS
17	2psw	4	1	P 1 21 1	0.452	0.107	543.2	-3.1	-0.000	HUMAN MAK3 HOMOLOG IN COMPLEX WITH COA
18	3tfy	4	2	P 1 21 1	0.448	0.100	545.6	-0.7	-0.000	NAA50P AMINO-TERMINAL ACETYLTRANSFERASE BOUND TO SUBSTRATE PEPTIDE FRAGMENT AND COA
19	1ro5	2	12	F 2 3	0.448	0.098	520.3	-4.8	0.119	CRYSTAL STRUCTURE OF THE AHL SYNTHASE LASI
20	3d8p	1	1	I 4 2 2	0.447	0.123	418.7	3.9	-0.000	CRYSTAL STRUCTURE OF ACETYLTRANSFERASE OF GNAT FAMILY (NP_373092.1) FROM STAPHYLOCOCCUS AUREUS MU50 AT 2.20 A RESOLUTION

Table 6.12 Interface similarity of the dimer in the crystal form $P2_12_12_1$

Examined 77708 entries, 1705706 interfaces
hits 1-20 of 167.

##	Entry	Intf No	mm Size	Space group	Q score	Seq. Id	Interface area, Å2	ΔG kcal/mol	CSS	Title
1	1xeb	3	2	P 32 2 1	0.797	0.352	931.5	-19.5	1.000	CRYSTAL STRUCTURE OF AN ACYL-COA N-ACYLTRANSFERASE FROM PSEUDOMONAS AERUGINOSA
2	3efa	1	2	P 65 2 2	0.634	0.210	840.3	-7.5	0.630	CRYSTAL STRUCTURE OF PUTATIVE N-ACETYLTRANSFERASE FROM LACTOBACILLUS PLANTARUM
3	1q2y	1	1	P 21 21 21	0.608	0.178	376.0	-0.5	0.000	CRYSTAL STRUCTURE OF THE PROTEIN YJCF FROM BACILLUS SUBTILIS: A MEMBER OF THE GCN5-RELATED N-ACETYLTRANSFERASE SUPERFAMILY FOLD
4	3p2h	1	1	P 32 2 1	0.481	0.179	521.0	-3.3	-0.000	CRYSTAL STRUCTURE OF TOFI IN A TERNARY COMPLEX WITH AN INHIBITOR AND MTA
5	2jdd	1	1	C 1 2 1	0.481	0.151	758.8	-4.1	-0.000	GLYPHOSATE N-ACETYLTRANSFERASE BOUND TO ACETYL COA AND 3-PHOSPHOGLYCERATE
6	2cnm	3	6	H 3 2	0.479	0.164	543.6	-3.2	1.000	RIMI - RIBOSOMAL S18 N-ALPHA-PROTEIN ACETYLTRANSFERASE IN COMPLEX WITH A BISUBSTRATE INHIBITOR (CTERM-ARG-ARG-PHE- TYR-ARG-ALA-N-ALPHA-ACETYL-S-COA).
7	2cns	4	1	H 3 2	0.478	0.160	543.4	-4.4	0.033	RIMI - RIBOSOMAL S18 N-ALPHA-PROTEIN ACETYLTRANSFERASE IN COMPLEX WITH ACETYLCOA.
8	2jdc	1	1	C 1 2 1	0.474	0.154	857.8	-5.4	0.014	GLYPHOSATE N-ACETYLTRANSFERASE BOUND TO OXIDIZED COA AND SULFATE
9	3p2f	1	1	P 32 2 1	0.472	0.185	498.3	-3.1	0.000	CRYSTAL STRUCTURE OF TOFI IN AN APO FORM
10	2bsw	1	2	C 1 2 1	0.469	0.144	849.1	-4.7	0.134	CRYSTAL STRUCTURE OF A GLYPHOSATE-N-ACETYLTRANSFERASE OBTAINED BY DNA SHUFFLING.
11	2cnt	7	12	P 3 2 1	0.467	0.153	535.4	-4.3	0.594	RIMI - RIBOSOMAL S18 N-ALPHA-PROTEIN ACETYLTRANSFERASE IN COMPLEX WITH COENZYMEA.
12	3f8k	1	1	P 21 21 21	0.448	0.109	618.4	-7.2	-0.000	CRYSTAL STRUCTURE OF PROTEIN ACETYLTRANSFERASE (PAT) FROM SULFOLOBUS SOLFATARICUS
13	2i79	1	2	P 21 21 2	0.444	0.124	1482.1	-2.2	0.462	THE CRYSTAL STRUCTURE OF THE ACETYLTRANSFERASE OF GNAT FAMILY FROM STREPTOCOCCUS PNEUMONIAE
14	1cjw	2	1	C 2 2 21	0.444	0.120	549.7	-5.3	0.014	SEROTONIN N-ACETYLTRANSFERASE COMPLEXED WITH A BISUBSTRATE ANALOG
15	1vhs	1	2	C 1 2 1	0.443	0.117	1242.7	-2.4	1.000	CRYSTAL STRUCTURE OF A PUTATIVE PHOSPHINOTHRICIN N-ACETYLTRANSFERASE
16	3lod	1	1	P 61 2 2	0.443	0.136	428.4	-1.2	0.000	THE CRYSTAL STRUCTURE OF THE PUTATIVE ACYL-COA N-ACYLTRANSFERASE FROM KLEBSIELLA PNEUMONIAE SUBSP.PNEUMONIAE MGH 78578
17	1l0c	2	1	C 2 2 21	0.442	0.103	552.0	-4.2	-0.000	INVESTIGATION OF THE ROLES OF CATALYTIC RESIDUES IN SEROTONIN N-ACETYLTRANSFERASE
18	1ro5	2	12	F 2 3	0.442	0.104	520.3	-4.8	0.119	CRYSTAL STRUCTURE OF THE AHL SYNTHASE LASI
19	3d8p	1	1	I 4 2 2	0.437	0.125	418.7	3.9	-0.000	CRYSTAL STRUCTURE OF ACETYLTRANSFERASE OF GNAT FAMILY (NP_373092.1) FROM STAPHYLOCOCCUS AUREUS MU50 AT 2.20 A RESOLUTION
20	2psw	4	1	P 1 21 1	0.436	0.109	543.2	-3.1	-0.000	HUMAN MAK3 HOMOLOG IN COMPLEX WITH COA

The top hit from this analysis was the structure 1XEB, an acyl coA-N-acetyltransferase from *P. aeruginosa* (Figure 6.8)

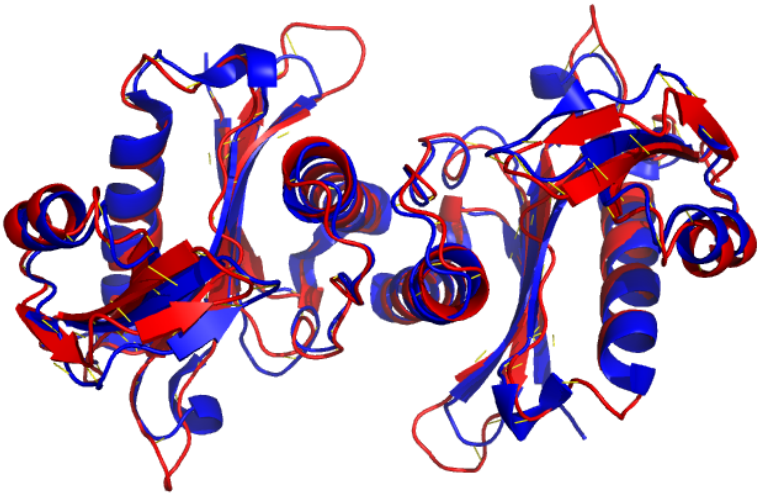


Figure 6.8 A similar dimer interface to BPSL0606: Superposition of the dimers in BPSL0606 (red) and an acyl coA-N-acetyltransferase from *P. aeruginosa* (PDB ID: 1XEB) (blue) shows their dimer interface similarity. This figure was produced using Pymol (Schrodinger, 2010).

The second hit was with the structure 3EFA, a putative N-acetyltransferase from *Lactobacillus plantarum*. This protein also forms a dimer using the same secondary structure elements as BPSL0606, however the arrangement of these elements at the two fold is quite different, with the helix α_1 packing in a crossed fashion with its symmetry related mate for the other subunit (Figure 6.9). This perhaps indicates that this surface can form more than one type of interaction. The proteins further down the list; an uncharacterised GCN5-related N-acetyltransferase from *B. subtilis* (PDB ID: 1Q2Y), an N-acyl-L-homoserine lactone synthase TofI from *B. glumae* (PDB ID: 3P2H) and a glyphosate N-acetyltransferase from *B. licheniformis* (PDB ID: 2JDD), all share the same subunit structure as BPSL0606 but appear to be monomers in their respective crystal structures. The ribosomal S18 N-α-protein acetyltransferase from *S. typhimurium* (PDB ID: 2CNM) forms a trimer in the crystal (Figure 6.10), indicating that different oligomeric assemblies are possible with this protein fold.

The top hits in this analysis are all members of the GCN5-related N-acetyltransferase (GNAT) superfamily, other members in this family occur as dimers, but using totally different interface to that seen in BPSL0606, e.g. aminoglycoside 6'N-acetyltransferases from *S. enteritidis* (PDB ID: 1S3Z) and from *Acinetobacter haemolyticus* (PDB ID: 4EVY) (Figure 6.11). As BPSL0606 shares the same subunit fold as this family, it may also be a GNAT superfamily member. This is discussed in the following chapter.

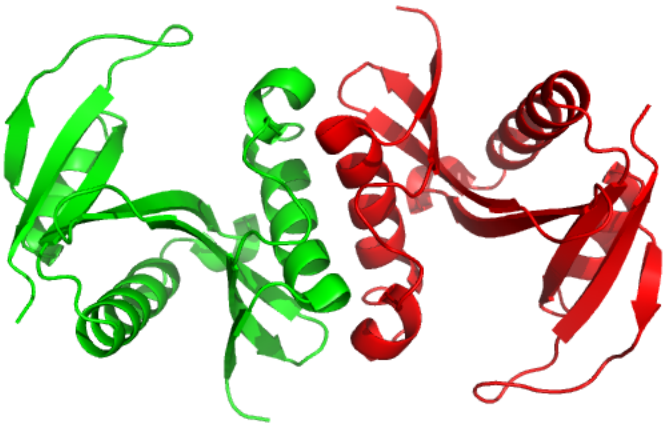


Figure 6.9 A similar dimer interface to BPSL0606 but different orientation: A dimer of a putative N-acetyltransferase from *L. plantarum* (PDB ID: 3EFA) is formed by the helix α_1, crossed with its symmetry mate. This figure was produced using Pymol (Schrodinger, 2010).

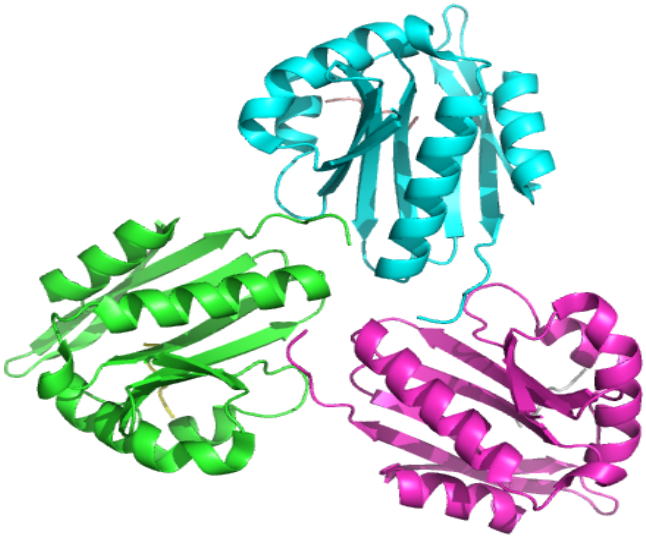
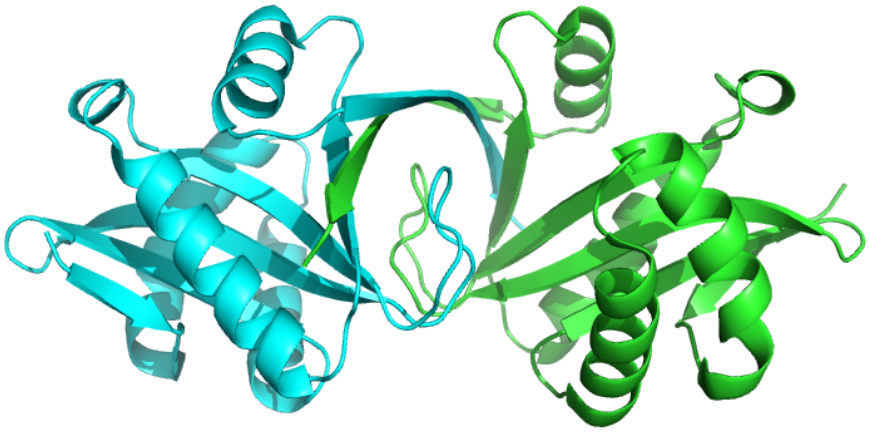
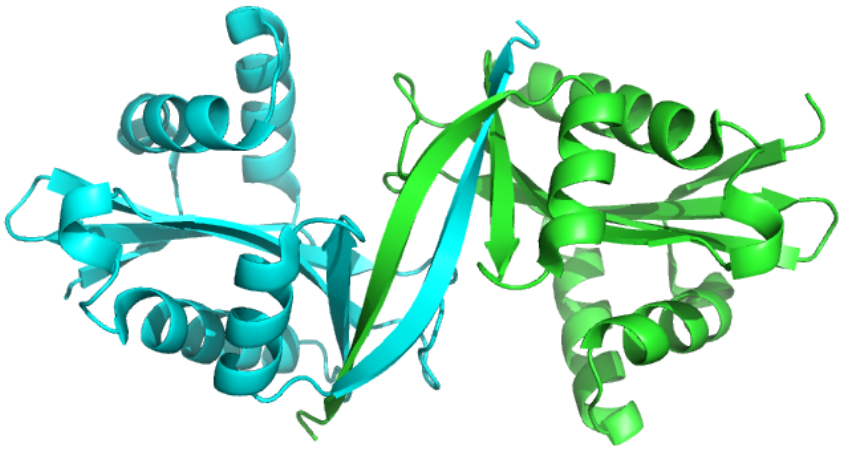


Figure 6.10 A different oligomeric assemblies with similar interface to BPSL0606: A ribosomal S18 N-α-protein acetyltransferase from *S. typhimurium* (PDB ID: 2CNM) shares a similar subunit structure to BPSL0606, but forms a trimer in the crystal. This figure was produced using Pymol (Schrodinger, 2010).



a)



b)

Figure 6.11 Different dimer interfaces to BPSL0606: Most of the GNAT members form dimers, some have different interfaces to BPSL0606; aminoglycoside 6'N-acetyltransferases from *S. enteritidis* (PDB ID: 1S3Z) (a) and from *A. haemolyticus* (PDB ID: 4EVY) (b). This figure was produced using Pymol (Schrodinger, 2010).

Chapter 7: Structural and functional analysis of BPSL0606

This chapter describes the comparison between the BPSL0606 structure and other similar known structures. A structure-based function prediction of BPSL0606 is included.

7.1. Structure similarity

As the structure of proteins is more conserved than the sequence, identifying a specific fold in a novel structure could assist a structure-based function prediction. However, the similarity of proteins at the structural level does not directly imply the protein function (Lee *et al*, 2007). This knowledge-based approach, by comparing the protein fold and structural motifs to those that have been characterised in the structural databases, is the most important method for predicting function from structure. Any structural similarity seen may infer possible biochemical and functional information for an uncharacterised protein (Thornton *et al*, 2000).

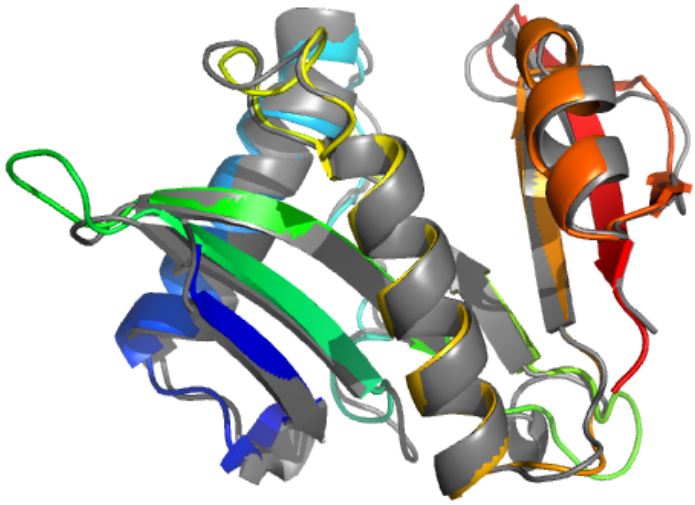
The coordinates of the BPSL0606 structure were compared against those available published in the protein data bank (PDB) using the Dali server (Holm L, 2010). The results from this analysis showed that BPSL0606 was structurally similar to several proteins related to N-acetyltransferases, including members of GCN5-related N-acetyltransferase superfamily (GNAT). The top ten similar structures to BPSL0606 are shown in Table 7.1. The structural similarity between BPSL0606 and the GNAT family may provide clues to the characteristics of BPSL0606.

At the protein sequence level, analysis of BPSL0606 using BlastP showed the sequence similarity to an ElaA family protein (a predicted transferase with an acyl-coA N-acyltransferase domain) of *Burkholderia* species and an N-acetyltransferase from other bacteria. For the protein structure level, the most

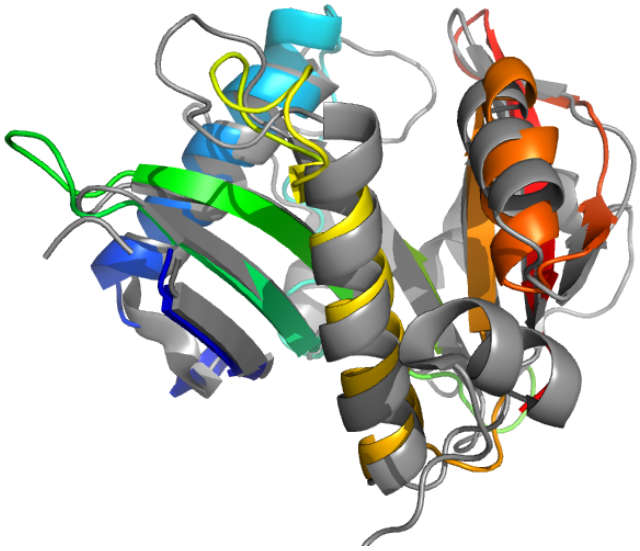
similar structure of characterised proteins to the BPSL0606 structure was the acyl homoserine lactone (AHL) synthase (LasI) from *P. aeruginosa*, which plays a role in virulence gene regulation in this opportunistic human pathogen (Gould *et al*, 2004). Like BPSL0606, the LasI structure contains the acyl coA N-acyltransferase fold as do members of the GNAT family.

Table 7.1 Top-ten hits for matching the BPSL0606 structure by Dali server

	PDB ID	**Z-score**	**RMSD (Cα)**	**% Sequence identity**	**No. of aligned/ total residues**	**Protein**	**Organism**
1	1XEB	24.6	1.0	35	144/149	Hypothetical protein	*P. aeruginosa*
2	3EFA	19.1	2.0	22	141/146	Putative acetyl transferase	*L. plantarum*
3	1Q2Y	18.2	2.0	18	136/140	Uncharacterised GCN5-related N acetyltransferase (GNAT)	*B. subtilis*
4	1RO5	15.6	2.5	10	145/197	Autoinducer synthetic protein (LasI)	*P. aeruginosa*
5	1CJW	15.6	2.6	9	140/166	Serotonin N-acetyltransferase	*Ovis aries* (sheep)
6	2JDD	15.4	2.4	14	133/145	Glyphosate N-acetyltransferase	*B. licheniformis*
7	1J4J	15.3	2.7	11	144/170	Tabtoxin resistance protein	*P. syringae pv. tabaci*
8	2CNM	15.2	2.3	13	134/151	Ribosomal s18 N-alpha-protein acetyltransferase (RimI)	*S. typhimurium*
9	1M4I	14.9	2.7	14	138/181	Aminoglycoside 2' acetyltransferase	*Mycobacterium tuberculosis*
10	2G0B	14.8	2.5	11	140/184	N-Acyl Amino Acid Synthase (FeeM)	uncultured bacterium



a)



b)

Figure 7.1 A superposition of the BPSL0606 structure (rainbow) and similar structures (grey): a) The most similar structure to BPSL0606 analysed by Dali was a hypothetical protein from *P. aeruginosa* (PDB ID 1XEB). b) The most similar structure with known function to BPSL0606 was an AHL synthase, LasI from *P. aeruginosa* (PDB ID 1RO5). This figure was produced using Pymol (Schrodinger, 2010).

7.2. Structural fold and motifs of GNAT family members

Since the BPSL0606 structure was structurally similar to a number of GNAT-fold containing proteins, the following section describes the structural fold and motifs of the GNAT family.

The GCN5 (general control non-derepressible 5) -related N-acetyltransferases (GNAT) belong to a superfamily of enzymes that catalyse the acetylation of various substrates by transferring the acetyl group of acetyl coenzyme A to a primary amine of the substrate (Dyda *et al*, 2000). The chemical reaction catalysed by GNATs is shown in Figure 7.2. Proteins from this family function in a wide variety of different biochemical processes, including the acetylation of aminoglycoside antibiotics (Vakulenko & Mobashery, 2003), eukaryotic chromatin modification by histone acetylation (Khan & Khan, 2010), the light-mediated regulation of melatonin production and mammalian circadian rhythm (Scheibner *et al*, 2002), N-α-acetylation of ribosomal proteins (Hu *et al*, 2010) and mycothiol biosynthesis (Vetting *et al*, 2006). Members of this superfamily that have been structurally characterised include aminoglycoside N-acetyltransferase (Burk *et al*, 2003; Vetting *et al*, 2002; Vetting *et al*, 2004; Vetting *et al*, 2008b; Wolf *et al*, 1998; Wybenga-Groot *et al*, 1999), serotonin N-acetyltransferase (Hickman *et al*, 1999; Scheibner *et al*, 2002; Wolf *et al*, 2002), histone N-acetyltransferase (Clements *et al*, 1999; Dutnall *et al*, 1998; Poux & Marmorstein, 2003; Rojas *et al*, 1999; Trievel *et al*, 1999), glucosamine-6-phosphate N-acetyltransferase (Peneff *et al*, 2001), mycothiol synthase (Vetting *et al*, 2003; Vetting *et al*, 2006), protein N-myristoyl-transferase (Bhatnagar *et al*, 1999; Brannigan *et al*, 2010), tabtoxin resistant protein (He *et al*, 2003), Rim N-α-acetyltransferase (Vetting *et al*, 2008a; Vetting *et al*, 2005a) and glyphosate N-acetyltransferase (Siehl *et al*, 2007).

The structures of these proteins have revealed the core GNAT fold (Figure 7.3), which consists of two antiparallel N-terminal β-strands (β0, β1) followed by two helices (α1, α2) and three further strands (β2-β4) that in turn, preceded a central helix (α3). The fold is completed by a further strand (β5), helix (α4) and a final strand (β6) (Vetting *et al*, 2005b). This common topology of the GNAT structural fold contains four conserved sequence motifs, called C, D, A and B, in the order

where the motifs follow each other in the primary sequences (Neuwald & Landsman, 1997).

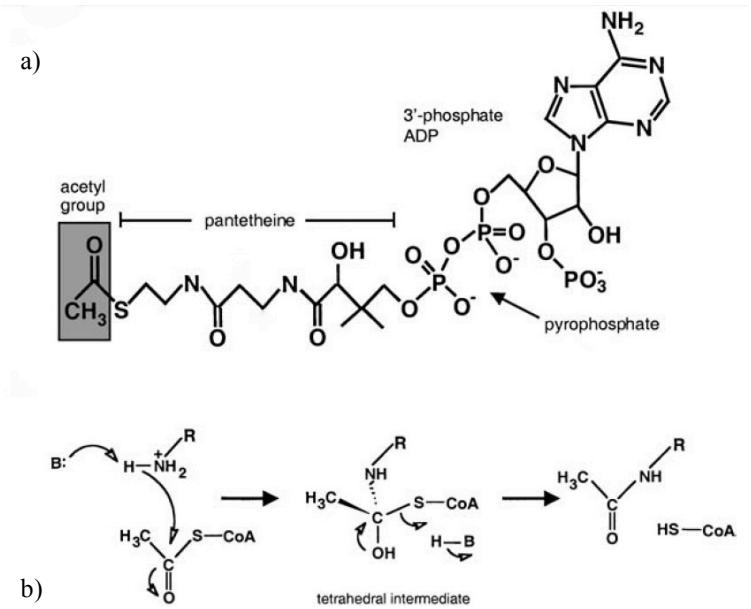


Figure 7.2 A schematic of the general reaction of GNATs catalysis: a) Acetyl coenzyme A is the acetyl group donor. b) The nucleophilic attack of a primary amine occurs on the acyl carbon of the acetyl group, resulting in a predicted tetrahedral intermediate. This figure was modified from Dyda *et al* (2000).

7.2.1. Motif C

The first motif from the N-terminus that has been identified is motif C, which has the weakest conservation at the primary sequence level within this superfamily compared to other motifs. The residues of this motif lie mostly on helix α1. However, as there is some variation in the secondary structure between members of the family in this area the motif is not strictly conserved. For example, in a histone N-acetyltransferase from yeast (yHat1) (Dutnall *et al*, 1998), the helix α1 is present but not the helix α2, which is present in other GNAT members. The absence of the second helix in the yHat1 structure is possibly accommodated by a shift in the position of the helix α1, compared to the analogue helix in other GNAT structures. Interestingly, in the case of a sheep serotonin N-acetyl transferase, the observations of the apo-enzyme structure and the complex form with a bisubstrate coenzyme A-S-acetyltryptamine analogue have shown conformational changes upon the acetyl coenzyme A binding in which the first and second helices were extended, facilitating the serotonin binding (Hickman *et al*, 1999).

7.2.2. Motifs D and A

Motif D lies on strands β2 and β3 and is adjacent to motif C in the three dimensional structure, and these two elements come together to form the first half of the molecule. After a short turn between the strands β3 and β4, the polypeptide sequence leads into motif A, which is the longest and universally conserved motif of this superfamily. This motif has a conserved sequence of Q/RXXGXGXXL. Motif A runs from strand β4 to the carboxyl end of the helix α3, which is the longest helix in the GNAT family. This motif is involved in acetyl coenzyme A binding and also plays an important role in catalysis. From the crystal structures of GNAT members in complex with acetyl coenzyme A, the acetyl coenzyme A is found in a V shaped cleft between the strands β4 and β5 (Sternglanz & Schindelin, 1999). Adjacent to the first turn of the helix α3, the pyrophosphate moiety of acetyl coenzyme A interacts with main-chain amides of the protein. This binding pocket is made by hydrogen bonds between amides of the N-terminal residues of the helix α3 and the phosphate oxygen atoms of the acetyl coenzyme A molecule. The pantothenate and β-mercaptoethanolamine parts of acetyl coenzyme A are placed along the strand β4 and stabilised by a number of hydrogen bonds (Figure 7.4).

7.2.3. Motif B

The final motif in the GNAT family is motif B, lying along strand β5 and ends at the C-terminal end of helix α4. This motif is also partly involved in coenzyme binding, with hydrophobic residues on helix α_4 forming a hydrophobic pocket that stabilises the bottom part of the acetyl coenzyme A binding site. There are some differences in the structures in this part of the fold between different family members, with, for example, a long extension between strand β5 and helix α4 in the structure of an aminoglycoside 6' N-acetyltransferase from *Enterococcus faecium* (Wybenga-Groot *et al*, 1999).

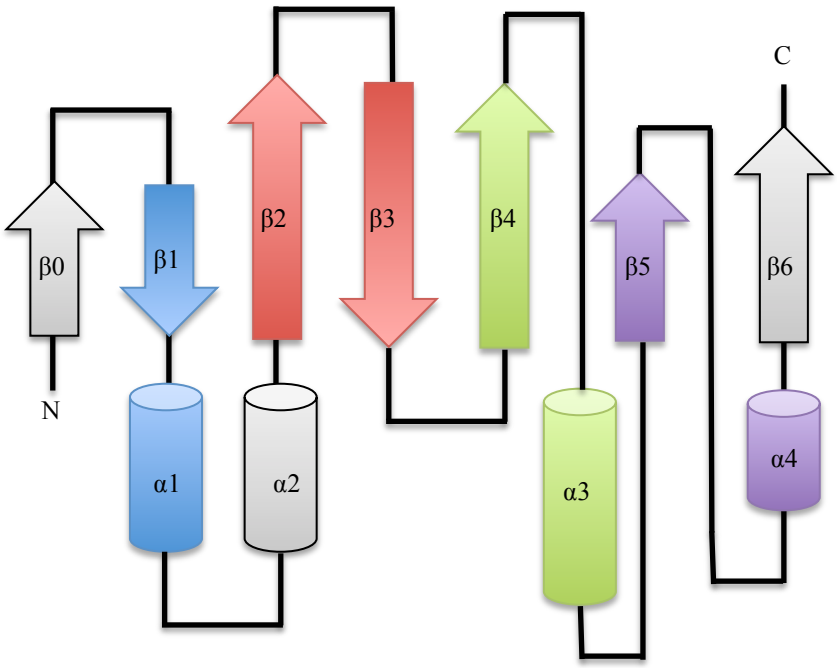


Figure 7.3 The topology of the GNAT fold represented by the secondary structure elements from the N-terminus: motif C-blue, motif D-red, motif A-green and motif B-purple.

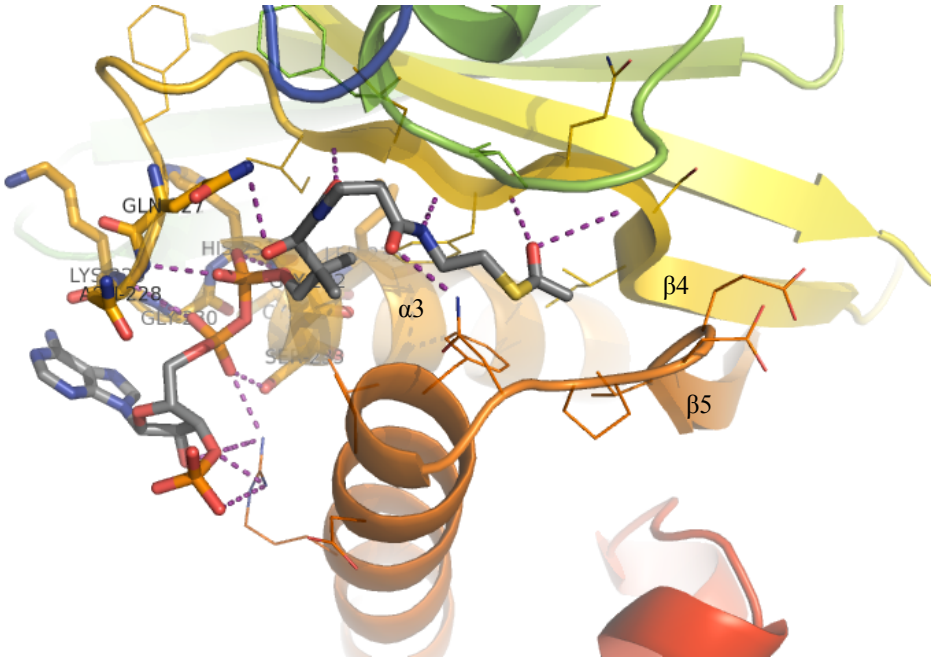


Figure 7.4 The acetyl coenzyme A binding site in the β_4-α_3-β_5 motif: The most conserved motif A (QXXGXGXXL) is shown in yellow sticks. Acetyl coenzyme A (grey stick) is bound in a histone N-acetyltransferase from yeast (PDB ID: 1BOB), with H-bonds (violet dashes). This figure was produced using Pymol (Schrodinger, 2010).

7.3. Structure comparison

In order to compare the structure of BPSL0606 to members of the GNAT superfamily, Cα superposition of the relevant structures were performed. The superposition of BPSL0606 and eight GNAT structures, including aminoglycoside N-acetyltransferases (PDB ID: 1B87 and 1N71), histone N-acetyltransferases (PDB ID: 1BOB, 1QSR, 1YGH and 1CM0) and serotonin N-acetyltransferases (PDB ID:1cjw) revealed that the backbone of BPSL0606 is similar to different members of this superfamily from different species (Figure 7.5). The sequence similarity seen in this family is shown in a structure based sequence alignment of BPSL0606 and selected GNAT members from different species (Figure 7.6).

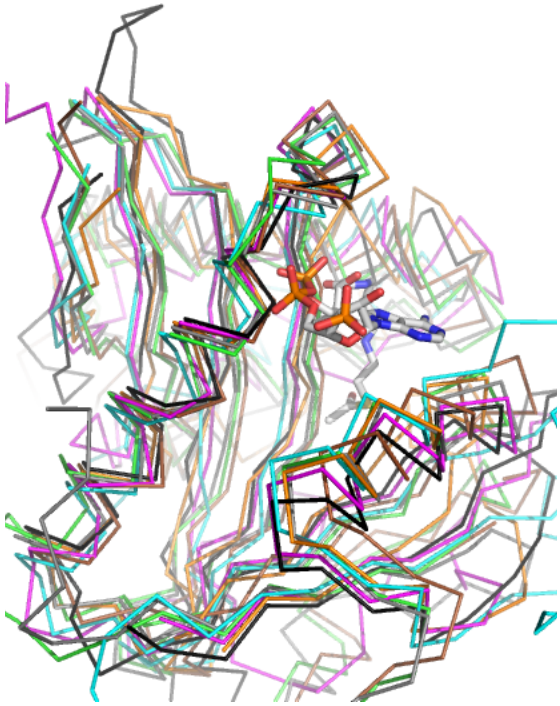


Figure 7.5 Superposition of the BPSL0606 structure and GNAT structures: The backbone of BPSL0606 colored in black is structurally similar to members of the GNAT superfamily (other colours), which have a unique fold responsible for acetyl coenzyme A binding (shown in stick). Overlapped structures included PDB ID: 1B87, 1N71, 1BOB, 1QSR, 1YGH, 1CM0 and 1CJW. This figure was produced using Pymol (Schrodinger, 2010).

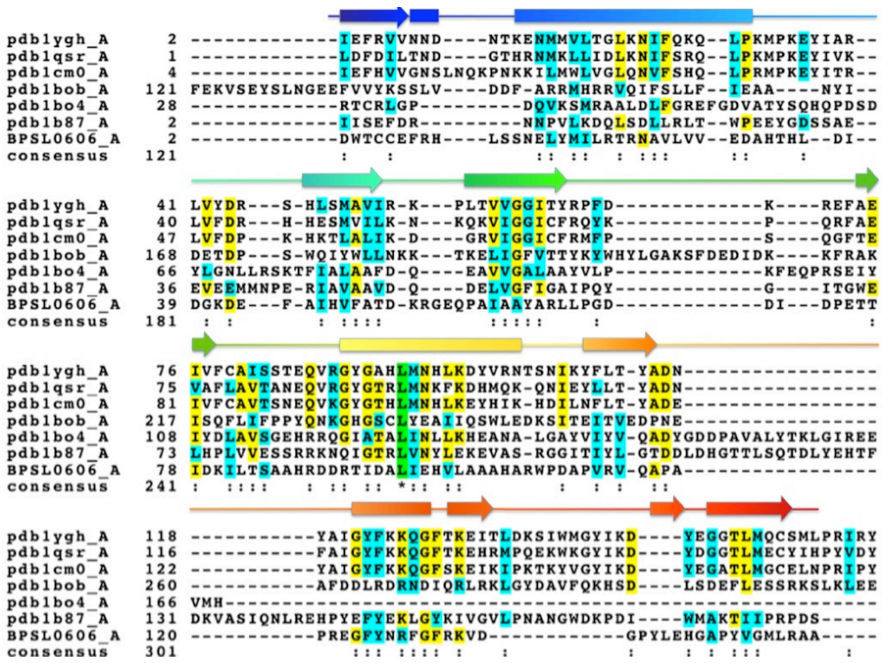


Figure 7.6 Structure-based alignment of BPSL0606 and GNAT proteins from different species. PDB ID: 1YGH (GCN5, yeast), 1QSR (GCN5, protozoa), 1CM0 (PCAF, human), 1BOB (HAT1, yeast), 1BO4 (AAC3', Gram negative bacteria) and 1B87 (AAC 6', Gram positive bacteria). Secondary structure elements representing the structure of BPSL0606 are shown from the N-terminus (blue) to the C-terminus (red). This figure was created using Boxshade (Hofmann & Baron).

To identify conserved residues within the GNAT superfamily, more GNAT structures that have been deposited in the protein data bank were taken into account. A structure based alignment of the known structures of GNATs and BPSL0606 (Figure 7.7) shows conserved residues in the known structures of proteins within the superfamily from different species. Using this alignment, conserved residues in the GNAT family were identified and plotted in the equivalent position in the structure of BPSL0606 (Figure 7.8). Within these conserved residues, most of those that are highly conserved are also present in BPSL0606. Differences between the GNATs conserved residues and the equivalent residues in BPSL0606 were inspected. Residue Ile 93 in BPSL0606 is different from the equivalent residue in the GNAT proteins, which is either Gly or Ala. As this residue lies under the pyrophosphate of acetyl coenzyme A in the binding pocket of the GNAT family, the increased size of isoleucine in the BPSL0606 structure may affect the ligand binding. Within the GNAT family, two hydrophobic residues form part of the hydrophobic pocket which lies underneath

```
PDB:1cm0:A          1 ----------------------KVIE-FHV-----V-GNSLNQKPNKKILMWLVGLQNVF
PDB:1cjw:A          1 -------------------HTLPANE-FRC-----L-TPE-----------DAAGVFEIE
PDB:1iid:A          1 ------------------MSLPDGFY-IRR-----M-EEG-----------DLEQVTETL
PDB:2c27:A        121 VGVRELIQMRRPLRDIPEPTIPDGVV-IRT-----Y-AGT----------SDDAELLRVN
PDB:2i79:A          1 --------------------MEYELL-IRE-----A-EPK-----------DAAELVAFL
PDB:2jdd:A          1 ------------------------IE-VKP-----I-N--------------AEDTYELR
PDB:1s3z:A          1 ----------------------SHMD-IRQ-----M-NKT-----------HLEHWRGLR
PDB:1n71:A          1 ------------------------MI-ISE-----FDRNNP---------VLKDQLSDLL
PDB:1s71:A          1 ----------------EIIPVSTTLE-LRA-----A-DES-----------HVPALHQLV
PDB:2cnm:A          1 ------------------------MNT-ISI-----L-STT-----------DLPAAWQIE
PDB:2prb:A          1 ----------------------SV-TLR-----L-MTE----------HDLAMLYEWL
PDB:1bob:A        105 EERKTHNLSDVFEKVSEYSLNGEEFV-VYK-----S-SLVD--------DFARRMHRRVQ
PDB:1b87:A          1 ------------------------MI-ISE-----FDRNNP---------VLKDQLSDLL
PDB:1qsr:A          1 ------------------------LD-FDI-----L-TNDGTH----RNMKLLIDLKNIF
PDB:2fiw:A          1 -----------------GHMVMSTPA-LRP-----Y-LPE-----------DAAVTAAIF
PDB:3exn:A          1 -------------------MHVLTLD-LAP-----V-TPK-----------DAPLLHRVF
PDB:3r1k:A          1 ------------------------VT-LCS-----P-TED-----------DWPGMFLLA
PDB:2vqy:A          1 ------------------------DSV-TLR-----L-MTE----------HDLAMLYEWL
PDB:1qsm:A          1 ----------------------DNIT-VRF-----V-TEN----------DKEGWQRLW
PDB:1j4j:A          1 ------------------------AQ-LRR-----V-TAESF-------AHYRHGLAQLL
PDB:1m4i:A          1 -------------------MHTQVHTARLV-----H-TADLD-------SETRQDIRQMV
BPSL0606.pdb:A      1 ------------------------MD-WTCCEFRHL-SSN-----------ELYMILRTR
consensus         121 ------------------------i--i-------v---------------d---l--v-

PDB:1cm0:A         46 ---LV-FDP-KH-----K--------------TLALIK--DG------RVIGGICFRM--
PDB:1cjw:A         40 ---EV-QH---F-----LTLC--------PELSLGWFV--EG------RLVAFIIGSLWD
PDB:1iid:A         38 ---FC-KLIKYW-----NEATVWNDKKIMQYNPMVIVD--KRT----ETVAATGNIII--
PDB:2c27:A        181 ---LA--E---R-----RGEAWF-----DPDGLILAFG--DSPRERPGRLLGFHWTKV--
PDB:2i79:A         43 SEEME-IF---LN---KQASSD-------NQITLLAFL--NG------KIAGIVNITADQ
PDB:2jdd:A         33 ---D---L---LR---GAF--------------HLGGYYG-G-------KLISIASFHQ--
PDB:1s3z:A         32 ---HL-ADGEEI-----LQAD--------HLASFIAMA--DG------VAIGFADASI--
PDB:1n71:A         36 ---EV-EE---M-----MNPE---------RIAVAAVD--QD------ELVGFIGAIPQY
PDB:1s71:A         53 ---VQ--G---N-----ILLHQR-----GYAKMYLIFC--QN------EMAGVLSFNAI-
PDB:2cnm:A         32 ---FF--G---N-----QGER----------YLNLKLTA--DD------RMAAFAITQV--
PDB:2prb:A         41 ---YLPSV---L-----AQES---------VTPYIAML--NG------EPIGYAQSYVAL
PDB:1bob:A        164 ---E---TDPSW-----Q--------------IYWLLNKKTK------ELIGFVTTYKYW
PDB:1b87:A         36 ---EV-EE---M-----MNPE---------RIAVAAVD--QD------ELVGFIGAIPQY
PDB:1qsr:A         40 ---LV-FDR-HH-----E--------------SMVILK--NK------QKVIGGICFRQ--
PDB:2fiw:A         55 ---FA--A---RL----SGQ-----------LTLIATL--QG------VPVGFASLKG--
PDB:3exn:A         46 ---DL-QT---L-----EVDP--------RRRAFLLFL--GQ------EPVGYLDAKL--
PDB:3r1k:A         31 ---A---TA-WR-----TLVP--------TDGAVVVRD--GS------EVVGMALYMD--
PDB:2vqy:A         43 ---YLPSV---LAQESVT--------------PYIAML--NG------EPIGYAQSYVAL
PDB:1qsm:A         36 ---DD---F---NFG--RFLDPN------IKMWAAVAVES-SS-----EKIIGMINFFN--
PDB:1j4j:A         50 ---L---KA--D-----IAAG--------SLLLWVVAE--DD------NVLASAQLSL--
PDB:1m4i:A         41 ---W---E---HT----LGG-----------MHALIWH--HG------AIIAHAAVIQ--
BPSL0606.pdb:A     40 ---G---K---D-----EFA-----------IHVFATD--KRGE--QPAIAAYARLLP--
consensus         241 ---------------------------------lva--------------vigfg------

PDB:1cm0:A         72 ---------------FPSQ-------GFTEIVFCAVT-SNEQVKGYGTHLMNHLKEYHIK
PDB:1cjw:A         72 EERLTQESLALHRPRGH----------SAHLHALAVH-RSFRQQGKGSVLLWRYLHHVGA
PDB:1iid:A         81 ---------------ERKIIH--ELGLCGHIEDIAVN-SKYQGQGLGKLLIDQLVTIGFD
PDB:2c27:A        219 ---------------HPDH------PGLGEVYVLGVD-PAAQRRGLGQMLTSIGIVSLAR
PDB:2i79:A         81 R-----------KRVRH----------IGDLFIVI-G-KRYWNNGLGSLLLEEAIEWAQA
PDB:2jdd:A         58 ---------------AEHSEL--QGGKQYQLRGMATL-EGYREQKAGSSLIKHAEEILRK
PDB:1s3z:A         65 ---------------RHDYVNGCDSSPVVFLEGIFVL-PSFRQRGVAKQLIAAVQRWGTN
PDB:1n71:A         67 ----------------GITP----------GWELHPLVVE-SSRRKNQIGTRLVNYLEKEVAS
PDB:1s71:A         86 ---------------EPIN----------KAAYIGYWLD-ESFQGQGIMSQSLQALMTHYAR
PDB:2cnm:A         60 ----------------VLD----------EATLFNIAVD-PDFQRRGLGRMLLEHLIDELET
PDB:2prb:A         73 GSGDG---WWEEETDPG-----------VRGIDQLLANASQLGKGLGTKLVRALVELLFN
PDB:1bob:A        193 HY---------IDKKFR-----------AKISQFLIF-PPYQNKGHGSCLYEAIIQSWLE
PDB:1b87:A         67 ---------------GIT----------GWELHPLVVE-SSRRKNQIGTRLVNYLEKEVAS
PDB:1qsr:A         67 ---------------YKPQ-------RFAEVAFLAVT-ANEQVRGYGTRLMNKFKDHMQK
PDB:2fiw:A         82 ----------------P------------DHIDMLYVH-PDYVGRDVGTTLIDALEKLAGA
PDB:3exn:A         76 ---------------GYPE------AEDATLSLLLIR-EDHQGRGLGRQALERFAAGLDG
PDB:3r1k:A         61 ---------------LRLTVPGEVVLPTAGLSFVAVA-PTHRRRGLLRAMCAELHRRIAD
PDB:2vqy:A         75 GSGDG---WWEEETDPG-----------VRGIDQLLANASQLGKGLGTKLVRALVELLFN
PDB:1qsm:A         72 ---------------HMTTWD---FKDKIYINDLYVD-ENSRVKGAGGKLIQFVYDEADK
PDB:1j4j:A         79 ---------------CQKPN----GLNRAEVQKLMVL-PSARGRGLGRQLMDEVEQVAVK
PDB:1m4i:A         67 ---------------RRLIYRG-NALRCGYVEGVAVR-ADWRGQRLVSALLDAVEQVMRG
BPSL0606.pdb:A     69 ---------------GDDI------DPETTIDKILTS-AAHRDDRTIDALIEHVLAAAHA
consensus         301 ----------------------------a-i--l-v-------kglgt-lm--l---v--

PDB:1cm0:A        109 HD-------IL-NFLT-YADE---Y-------------------------AIGYFKKQ-G
PDB:1cjw:A        121 QPA------VR-RAVL-MCED---A-------------------------LVPFYQRF-G
PDB:1iid:A        123 YG-------CY-KIIL-DCDE---K-------------------------NVKFYEKCGF
PDB:2c27:A        257 RLKTLDPAVEP-AVLL-YVES---DNV-----------------------AAVRTYQSL-G
PDB:2i79:A        118 SGI------LR-RLQL-TVQT---RNQ-----------------------AAVHLYQKH-G
PDB:2jdd:A        100 RG-------AD-LLWC-NART---S-------------------------ASGYYKKL-G
PDB:1s3z:A        109 KG-------CR-EMAS-DTSP---ENT-----------------------ISQKVHQALGF
PDB:1n71:A        102 RG-------GI-TIYL-GTDDLDHGTTLSQTDLYEHTFDKVASIQNLREHPYEFYEKL-G
PDB:1s71:A        122 RGD------IR-RFVI-KCRV---DNQ-----------------------ASNAVARRN-H
PDB:2cnm:A         95 RG-------VV-TLWL-EVRA---SNA-----------------------AAIALYESL-G
PDB:2prb:A        119 DPE------VT-KIQT-DPSP---SNL-----------------------RAIRCYEKAGF
PDB:1bob:A        232 DKS------IT-EITVEDPNE---A-------------------------FDDLRDRN-D
PDB:1b87:A        102 RG-------GI-TIYL-GTDDLDHGTTLSQTDLYEHTFDKVASIQNLREHPYEFYEKL-G
PDB:1qsr:A        104 QN-------IE-YLLT-YADN---F-------------------------AIGYFKKQ-G
PDB:2fiw:A        114 RG-------AL-ILTV-DASD---N-------------------------AAEFFAKR-G
PDB:3exn:A        114 V--------RRLYAV-VYGH---NP------------------------KAKAFFQAQGF
PDB:3r1k:A        105 SG-------YP-VAAL-HASE---GG------------------------GRF-G
PDB:2vqy:A        121 DPE------VT-KIQT-DPSP---SNL-----------------------RAIRCYEKAGF
PDB:1qsm:A        113 LG-------TP-SVYW-CTDE---SNH-----------------------RAQLLYVKV-G
PDB:1j4j:A        119 HK-------RG-LLHL-DTEA---GS------------------------VAEAFYSAL-A
PDB:1m4i:A        110 AY--------Q-LGAL-SSSA---R-------------------------ARRLYASR-G
BPSL0606.pdb:A    107 RWP------DA-PVRV-QAPA---P-------------------------REGFYNRF-G
consensus         361 -g-----------i-l-----------------------------------a--yy-k--g

PDB:1cm0:A        131 ------------------------------------F--SKEIKIPKTKYVGYIKDYEGA
PDB:1cjw:A        144 ------------------------------------F--HPAGPCAIVVGSLTFTEMHCS
PDB:1iid:A        146 ------------------------------------S--NAGVEMQIRK-----------
PDB:2c27:A        289 ------------------------------------F--TTYSVDTAYALA---------
PDB:2i79:A        144 ------------------------------------F--VIEGSQERGAYIEEGKFIDVY
PDB:2jdd:A        122 ------------------------------------F--SEQGEVFDTPPVGPHILMYKR
PDB:1s3z:A        135 ------------------------------------E--ETERVIFYRKRC---------
PDB:1n71:A        152 ------------------------------------Y--KIVGVLPNANGWDKPDIWMAK
PDB:1s71:A        148 ------------------------------------F--TLEGCMKQAEYLNGDYHDVNM
PDB:2cnm:A        120 ------------------------------------F--NEATIRRNYYPTAQGHEDAII
PDB:2prb:A        146 ------------------------------------E--RQGTVTTPDGPAVYMVQTRQA
PDB:1bob:A        256 IQRLRKLGYDAVFQKHSDLSDEFLESSRKSLKLEERQFNRLVEMLLLLNNS---------
PDB:1b87:A        152 ------------------------------------Y--KIVGVLPNANGWDKPDIWMAK
PDB:1qsr:A        126 ------------------------------------F--TKEHRMPQEKWKGYIKDYDGG
PDB:2fiw:A        136 ------------------------------------Y--VAKQRNTVSINGEWLANTTMT
PDB:3exn:A        138 RYVKDG-------------------------GPTLTWY--VRPL-----------------
PDB:3r1k:A        125 ------------------------------------Y--GPATTLHELTVDRRFARFHAD
PDB:2vqy:A        148 ------------------------------------E--RQGTVTTPDGPAVYMVQTRQA
PDB:1qsm:A        138 ------------------------------------Y--KAPKILYKRKGY---------
PDB:1j4j:A        143 ------------------------------------Y--TRVGELPGYCATPDGRLHPTA
PDB:1m4i:A        131 ------------------------------------W--LPWHGPTSVLAPTGPVRTPDD
BPSL0606.pdb:A    130 ------------------------------------F--RKVDGPYLEHGAPYVGMLRAA
consensus         421 ------------------------------------f-------i---------------
```

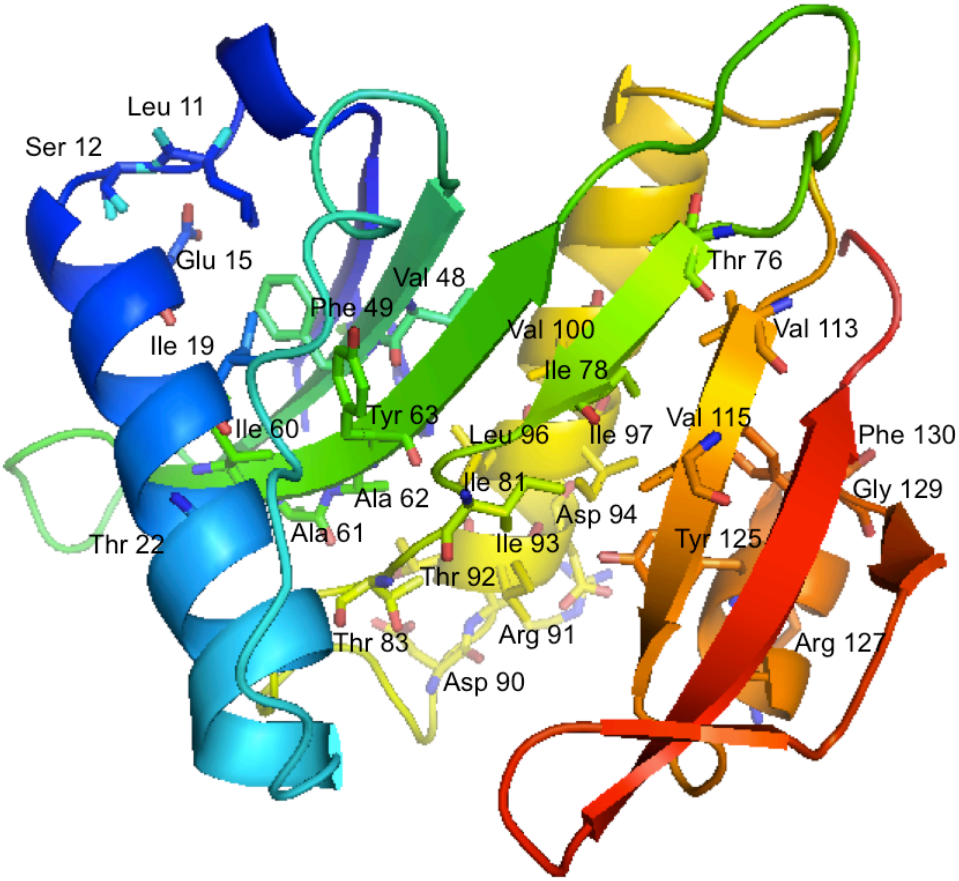
Motif C

Motif D

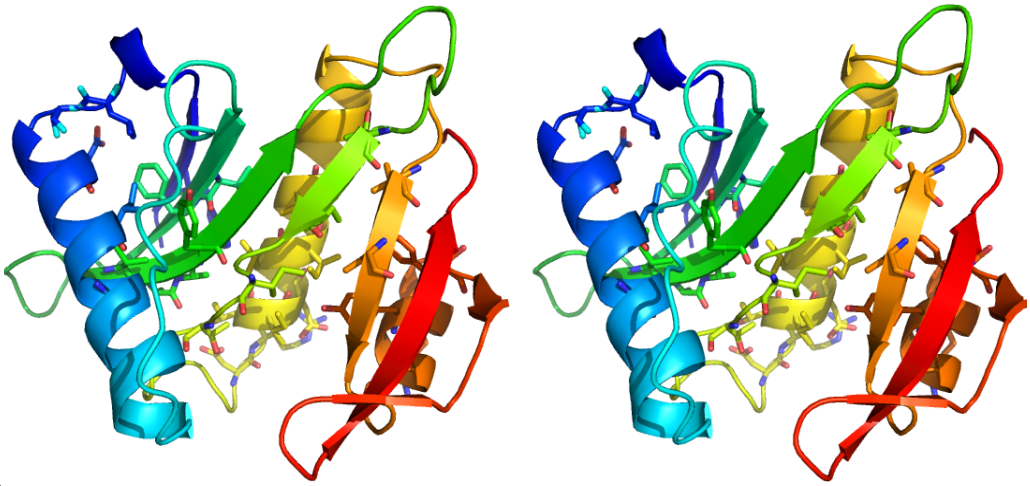
Motif A

Motif B

Figure 7.7 Structure based sequence alignment among the GNAT members and BPSL0606. Conserved residues among the known structures of GNATs were identified in box shades. Identical residues were highlighted in yellow and similar residues were highlighted in cyan. This figure was created using Boxshade (Hofmann & Baron).



a)



b)

Figure 7.8 Conserved residues of BPSL0606: a) Equivalent residues that are conserved within the GNAT superfamily are plotted on the structure of BPSL0606 (N-terminus-blue, C-terminus-red). b) Stereo representation of a). This figure was produced using Pymol (Schrodinger, 2010).

the pantotheine arm of acetyl coenzyme A. There is some variability in the residues at these two positions in the family, but the two residues seen in BPSL0606, Ile 81 and Tyr 125 are both hydrophobic and could provide the same function in this protein.

In addition to the analysis described above, the BPSL0606 structure was analysed using the Cathedral server (Redfern *et al*, 2007) for protein structure classification. As expected, there were significant structural matches between the BPSL0606 structure and the GNAT superfamily members, which have been classified into the superfamily 3.40.630.30. According to the CATH database, the superfamily 3.40.630.30 is in a class α-β (3) with architecture of 3-layer sandwich (40) containing a topology fold of aminopeptidases (630) and unassigned homology (30). Transferases form the majority of the members of this superfamily (90%) including acyltransferases that transfer groups other than amino-acyl groups (EC 2.3.1.xx). The GNAT superfamily has been also categorised in this category.

The structural and sequence comparison of BPSL0606 gave strong indications that this protein may well have an acetyl transferase activity, perhaps catalyzing the transfer of an acetyl group from acetyl coenzyme A to the primary amine of the substrate. The next section of this chapter describes attempts to show the binding of acetyl coenzyme A to BPSL0606.

7.4. Cocrystallisation of BPSL0606 and acetyl co-enzyme A

Attempts to crystallise BPSL0606 with acetyl coenzyme A were made. Different concentrations of acetyl coenzyme A solution were mixed with freshly purified BPSL0606, prior to crystallisation trials. Inspection of the conditions that have been used to prepare complexes of various GNAT members with acetyl coenzyme A or coenzyme A (Table 7.2), the acetyl coenzyme A concentration used in the experiments with BPSL0606 ranged over 0.5 mM, 1 mM and 2 mM, which were equivalent to 1-fold, 2-fold and 4-fold molar excess, respectively. A mixture of 10 mg/ml purified BPSL0606 protein and 2 mM acetyl coenzyme A solution was prepared. The hanging-drop based crystallisation was manually set up using a reservoir solution, which consisted of 14-24% PEG 3350, 0.1 M Bis-Tris propane

pH 6.3 and 0.2 M potassium thiocyanate, with 1:1, 1:2 and 2:1 ratios of protein mixture and reservoir solution.

Table 7.2 Crystallisation conditions for a number of GNAT proteins.

PDB ID	GNAT proteins	Organism	Crystallisation conditions		References
			Protein concentration (mg/ml)	Acetyl coenzyme A concentration	
1QSR	GCN5 related proteins	*Tetrahymena*	7.5	1.5 mM	Rojas *et al* (1999)
1B87	Aminoglycoside 6' N-acetyltransferase	*E. faecium*	7-10	2 mM	Wybenga-Groot *et al* (1999)
1BO4	Aminoglycoside 3' N-acetyltransferase	*Serratia marcescens*	3.5	5 mM (coenzyme A)	Wolf *et al* (1998)
1BOB	Histone N-acetyltransferase	*Saccharomyces cerevisiae*	10	2-fold molar excess	Dutnall *et al* (1998)
1CM0	PCAF (p300/CBP-associating factor)	Human	10	2-fold molar excess (sodium acetyl coenzyme A)	Clements *et al* (1999)

X-ray data sets from three crystals, one with 0.5 mM acetyl coenzyme A and two with 2 mM acetyl coenzyme A, were collected using the Diamond synchrotron (Table 7.3). Data processing was in similar manner as mentioned before. The structure determination was performed using molecular replacement in Phenix (Adams et al, 2010) and the solutions were obtained with LLG 2688.8, 3751.4 and 3467.5 for the BPSL0606 crystals with 0.5 mM acetyl coenzyme A, and BPSL0606 with 2 mM acetyl coenzyme A, respectively.

Models for each structure were rebuilt in Coot and the electron density maps were inspected for positive density features in the putative acetyl coenzyme A binding site. No electron density was present for acetyl coenzyme A in all three crystals (Figure 7.9). The lack of any density for the coenzyme may be because BPSL0606 actually does not bind acetyl coenzyme A, or that an intermediate is required for binding, or that the crystallisation conditions are not favourable for complex formation.

Table 7.3 Data collection statistics of crystals from the co-crystallisation experiments

Data set	**BPSL0606 +0.5 mM AcCoA**	**BPSL0606 +2 mM AcCoA**	**BPSL0606 +2mM AcCoA**
Spacegroup	P 1 2_1 1	P 1 2_1 1	P 1 2_1 1
Unit cell parameters:			
a (Å)	60.88	60.51	60.71
b (Å)	81.31	81.08	80.55
c (Å)	78.46	77.88	77.73
α (°)	90.00	90.00	90.00
β (°)	99.14	99.57	98.68
γ (°)	90.00	90.00	90.00
Temperature (K)	100	100	100
X-ray Source	DIAMOND I24	DIAMOND I03	DIAMOND I03
Detector		ADSC Q315 CCD	ADSC Q315 CCD
Resolution (Å)[1]	36.00-2.81 (2.89-2.81)	55.76-2.18 (2.24-2.18)	31.97-2.31 (2.37-2.31)
Energy (keV)	(0.97700)	12.700	12.700
Unique observations[1]	18136 (1342)	36560 (2759)	32423 (2392)
R_{merge} [1,2]	0.114 (0.897)	0.074 (0.462)	0.082 (0.735)
R_{pim} [1]	0.092 (0.559)	0.099 (0.433)	0.063 (0.545)
Completeness (%)[1]	98.1 (98.7)	94.6 (96.3)	98.9 (98.9)
Multiplicity[1]	2.5 (2.6)	2.3 (2.4)	3.3 (3.3)
Mean((I)/sd(I))[1]	7.2 (2.0)	8.8 (2.2)	7.8 (1.8)

[1] Numbers in parentheses indicate values for the highest resolution shell

[2] $R_{merge} = \Sigma_{hkl} \Sigma_i |I_i(hkl) - \langle I(hkl)\rangle| / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where $\langle I(hkl)\rangle$ is the mean intensity of the reflection

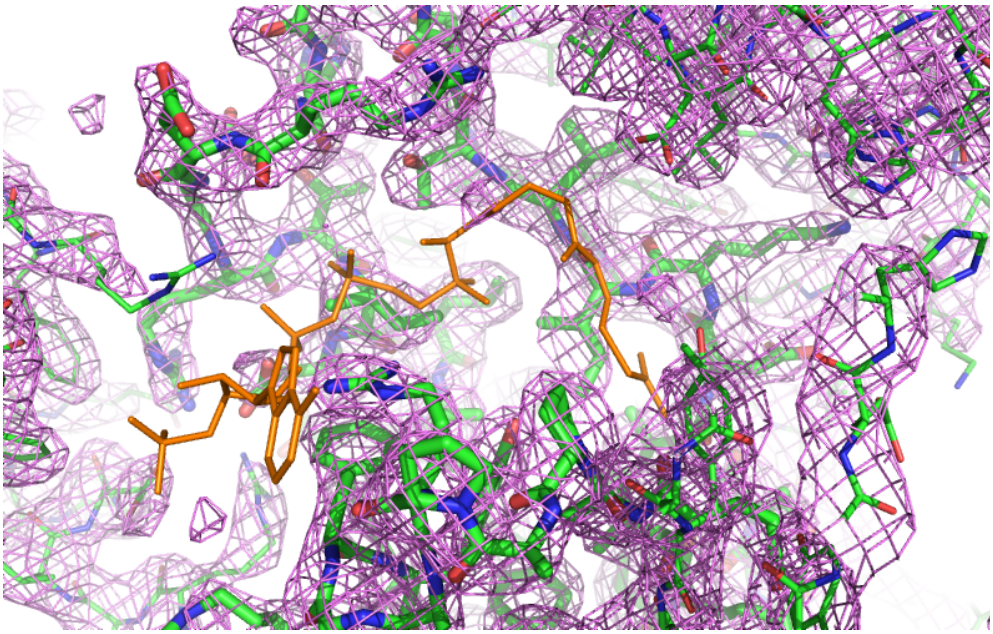


Figure 7.9 No electron density for acetyl coenzyme A: The electron density map, contoured at 1σ, around the putative binding site of acetyl coenzyme A in BPSL0606 is shown in violet. The modeled position of acetyl coenzyme A seen in other GNAT superfamily members is shown in orange, indicating that acetyl coenzyme A did not bind to BPSL0606. This figure was produced using Pymol (Schrodinger, 2010).

Further soaking experiments were carried out in order to increase the amount of acetyl coenzyme A and add magnesium chloride, which could provide a divalent atom to stabilise the charge on the pyrophosphate of the cofactor, which may be required for ligand binding. For crystal soaking, BPSL0606 crystals were mounted by a loop and transferred into a drop of solution containing 25% ethylene glycol, 22% PEG3350, 0.1 M Bis-Tris Propane pH 6.3, 0.2 M potassium thiocyanate, 10 mM magnesium chloride and 2 or 10 mM acetyl coenzyme A. The crystals were removed from the solution after soaking for 30-60 minutes.

X-ray data collection from five crystals, which had been soaked with acetyl coenzyme A and magnesium chloride were performed. Statistics for the data collection are shown in Table 7.4. Solutions from molecular replacement were rebuilt and inspected in Coot. Yet again, there was no electron density for acetyl coenzyme A (Figure 7.10).

Table 7.4 Data collection statistics of crystals from the soaking experiments

Data set	**BPSL0606 + 10 mM AcCoA 1 hr**	**BPSL0606 + 2 mM AcCoA 1 hr**	**BPSL0606 + 10 mM AcCoA 1 hr**	**BPSL0606 + 10 mM AcCoA O/N**	**BPSL0606 + 10 mM AcCoA O/N**
Spacegroup	P 1 2_1 1	P 1 2_1 1	P 1 2_1 1	P 1 2_1 1	P 1 2_1 1
Unit cell parameters					
a (Å)	60.40	60.62	60.54	60.49	60.70
b (Å)	80.51	80.65	80.38	80.75	81.06
c (Å)	77.48	77.96	77.47	78.11	78.53
α (°)	90.00	90.00	90.00	90.00	90.00
β (°)	98.71	98.79	98.74	98.93	99.23
γ (°)	90.00	90.00	90.00	90.00	90.00
Temperature (K)	100	100	100	100	100
X-ray Source	DIAMOND I03	DIAMOND I03	DIAMOND I03	DIAMOND I03	DIAMOND I03
Detector	ADSC Q315 CCD	ADSC Q315 CCD	ADSC Q315 CCD	ADSC Q315 CCD	ADSC Q315 CCD
Resolution (Å)[1]	25.53-2.31 (2.37-2.31)	26.58-2.25 (2.30-2.25)	44.01-1.89 (1.94-1.89)	33.45-2.04 (2.09-2.04)	51.58-1.99 (2.05-1.99)
Energy (keV)	12.700	12.700	12.700	12.700	12.700
Unique observations[1]	32176 (2390)	34516 (2519)	58283 (4245)	45900 (3421)	50747 (3739)
R_{merge} [1,2]	0.082 (0.420)	0.067 (0.656)	0.042 (0.593)	0.051 (0.406)	0.044 (0.550)
R_{pim} [1]	0.066 (0.33)	0.061 (0.469)	0.035 (0.446)	0.051 (0.399)	0.036 (0.412)
Completeness (%)[1]	99.0 (99.8)	97.4 (97.5)	99.0 (98.0)	97.1 (97.5)	99.0 (99.3)
Multiplicity[1]	3.3 (3.4)	3.4 (3.6)	3.4 (3.3)	2.7 (2.6)	3.3 (3.4)
Mean((I)/sd(I))[1]	8.6 (2.3)	11.2 (2.0)	14.0 (2.1)	11.7 (2.3)	13.7 (2.2)

[1] Numbers in parentheses indicate values for the highest resolution shell

[2] $R_{merge} = \Sigma_{hkl} \Sigma_i |I_i(hkl) - \langle I(hkl)\rangle| / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where $\langle I(hkl)\rangle$ is the mean intensity of the reflection

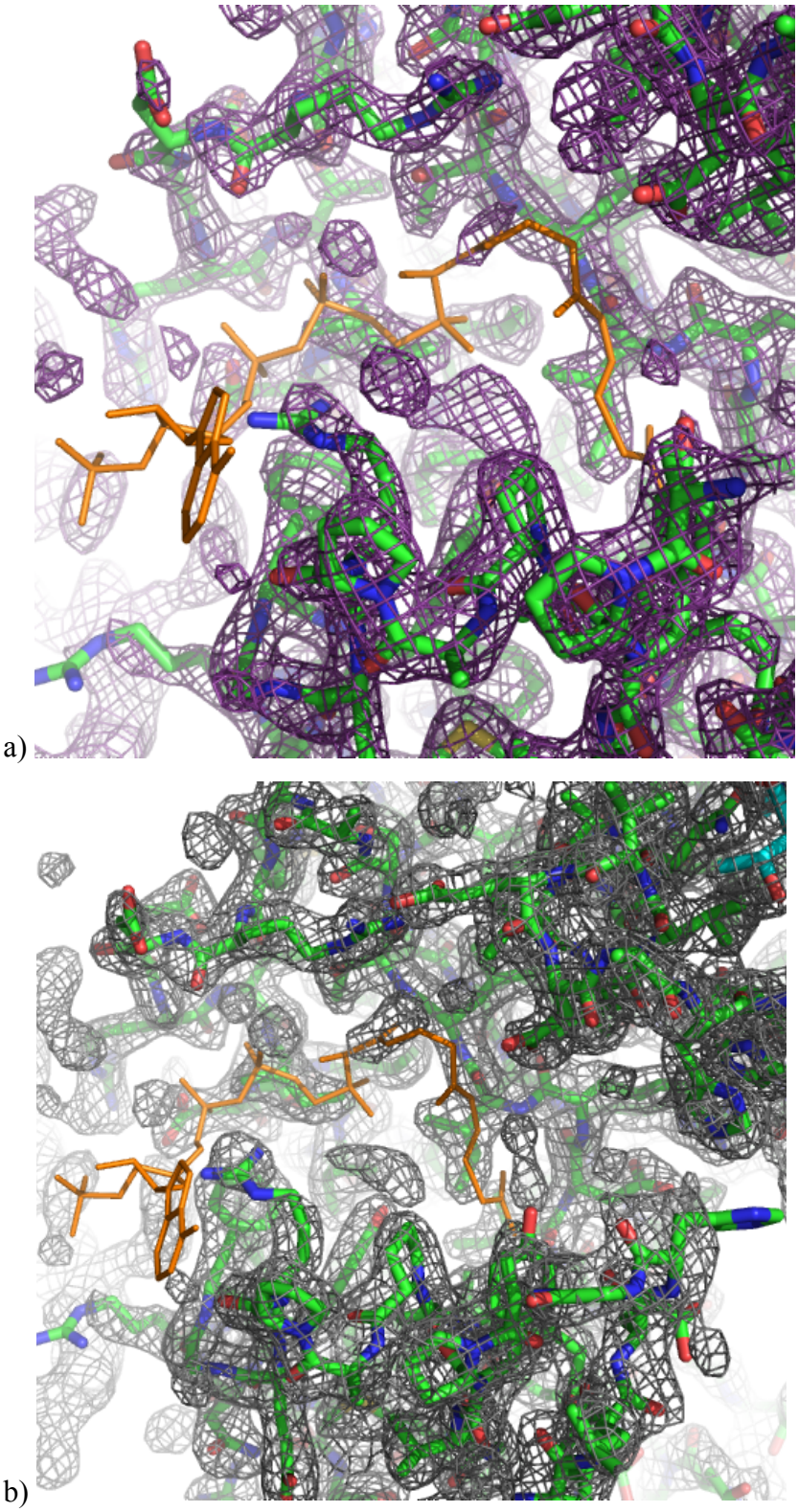


Figure 7.10 No acetyl coenzyme A bound to BPSL0606: Electron density maps, contoured at 1σ, derived from BPSL0606 crystals soaking with 10 mM acetyl coenzyme A and 10 mM magnesium chloride for 1 hour (a) and over-night (b). These results indicated that the acetyl coenzyme A did not bind to BPSL0606. This figure was produced using Pymol (Schrodinger, 2010).

7.5. Ligand observation by NMR

One way to identify protein-ligand interactions is ligand observation using the nuclear magnetic resonance (NMR) technique (Jahnke & Widmer, 2004). As the co-crystallisation and soaking experiments had failed to show any acetyl coenzyme A binding to BPSL0606, it was decided to use this method in an attempt to show acetyl coenzyme A binding. The principle of this technique lies on the detectable changes in atomic resonance properties of the ligand when it is bound to a protein and the free ligand molecule. Upon ligand binding, a bound ligand temporarily behaves like a large molecule. Unlike the unbound form, acting as a small molecule, slower tumbling in the bound ligand is observed and this behaviour results in the different NMR properties including faster relaxation, slower diffusion and large and negative nuclear Overhauser effects (NOEs). These differences can be detected by measuring ^{1}H NMR of ligand spectra with and without the target protein. Signals of a small molecule like ligand free in solution exhibit sharp line widths (Figure 7.11) whereas the broad line-width spectrum is observed if the ligand is bound to the target protein (Stockman & Dalvit, 2002).

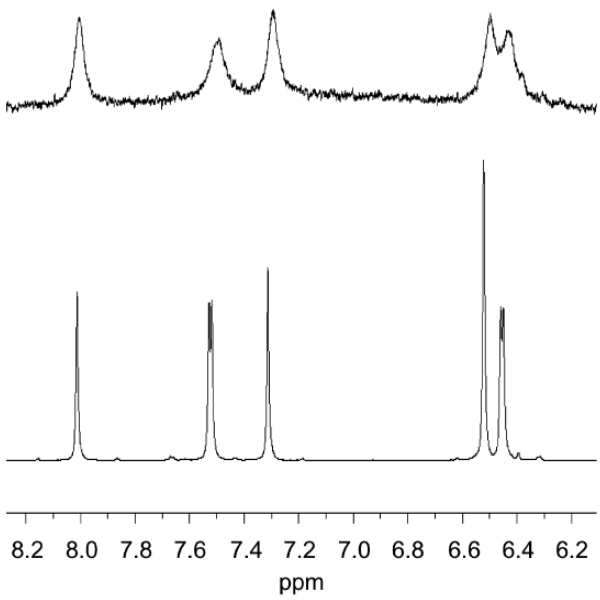


Figure 7.11 An example of ligand binding observation: Top is an expanded region of 1D 800 MHz ^{1}H NMR spectrum of a small molecule bound to the protein, which exhibits broad line-widths as seen on protein binding. The lower trace shows the sharp line-width NMR spectrum of the free ligand. This illustration was taken from Stock and Dalvit (2002).

To investigate the binding of BPSL0606 and acetyl coenzyme A, a 1:1 ratio mixture of BPSL0606 and acetyl coenzyme A solution was observed using ^{1}H NMR. Prior to the experiment, heavy water (D_2O) was added to the sample in order to prevent the interference of water solvent signal, resulting in the sharp NMR signal of the sample. The results showed that the behavior of the ligand-protein mixture in comparison to the ligand itself profile did not change (Figure 7.12). If the acetyl coenzyme A binds to BPSL0606, a broad line width spectrum should be recorded. The spectrum difference confirmed that no significant broad line widths were detected (Figure 7.13), indicating that BPSL0606 did not bind to acetyl coenzyme A. Alternatively, if the binding and release of the coenzyme occurred at a very fast rate, then this technique would not be able to detect any such binding.

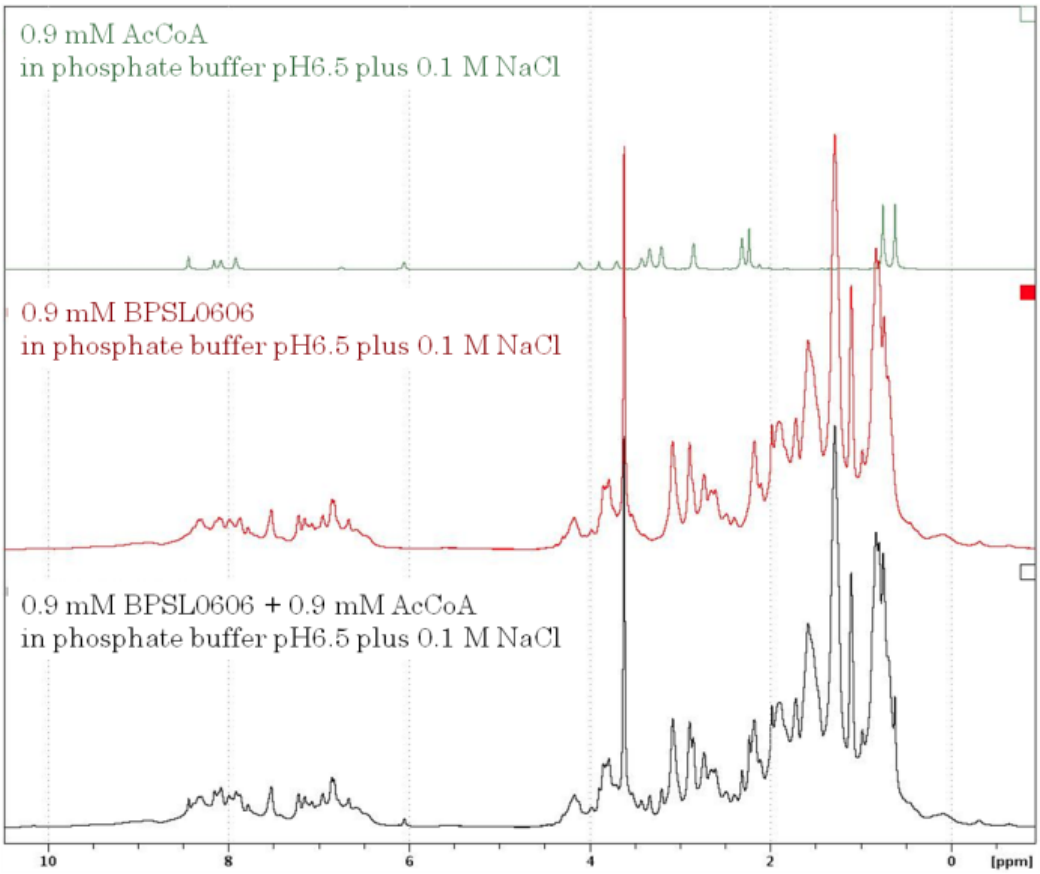


Figure 7.12 1D ^{1}H NMR spectra at 600 MHz for ligand observation: ^{1}H-NMR spectrum of the 1:1 molar ratio of BPSL0606 and acetyl coenzyme A mixture (black) shows no significant signal for binding recorded.

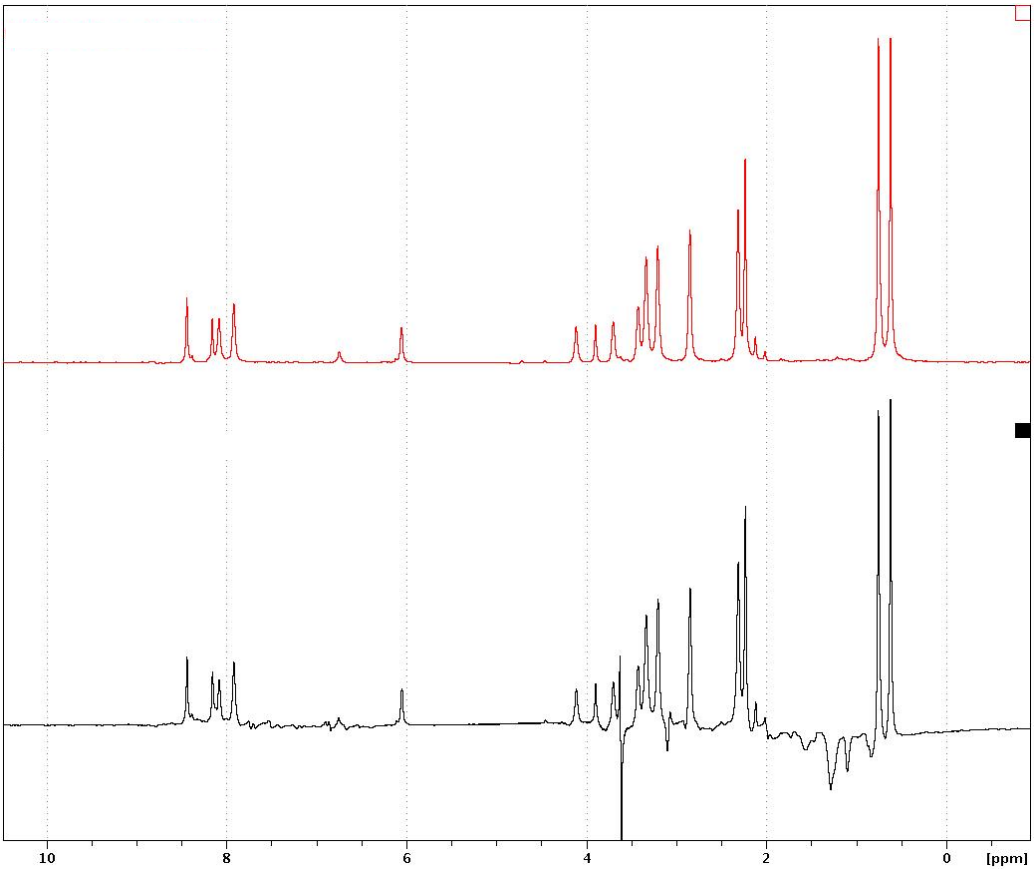


Figure 7.13 1D ^{1}H NMR spectra at 600 MHz of the acetyl coenzyme A solution: Top is an expanded region of the reference spectrum of acetyl coenzyme A in the absence of BPSL0606. Bottom is the difference of the spectrum of the mixture of acetyl coenzyme A and BPSL0606 when the protein spectrum was subtracted. No signal is indicated for the ligand binding.

7.6. Analysis of acetyl coenzyme A binding site

As no direct evidence could be obtained for acetyl coenzyme A binding to BPSL0606, an analysis of the coenzyme binding site in the GNAT family was undertaken, to see of it would be possible to model acetyl coenzyme A binding to BPSL0606. A signature feature of the GNAT superfamily is a β-bulge between β-strands β4 and β5, where the acetyl coenzyme A or coenzyme A is bound the proteins (Dyda *et al*, 2000). Binary and ternary complex structures of GNAT proteins and acetyl coenzyme A/ coenzyme A have revealed the acetyl coenzyme A binding site at this opening V-cleft (Figure 7.14).

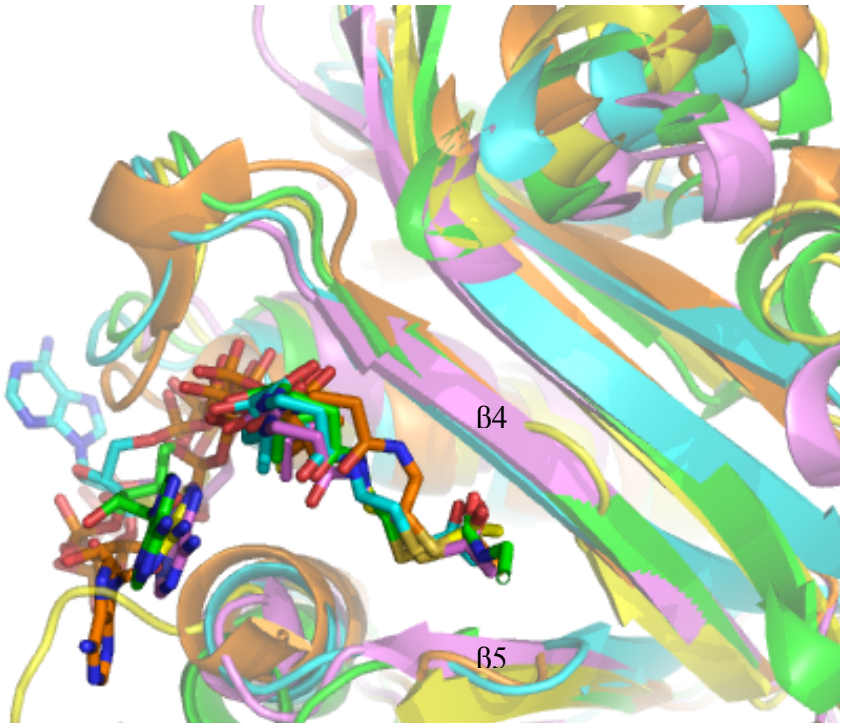


Figure 7.14 Superposition of the acetyl coenzyme A bound GNAT structures: The acetyl coenzyme A binding site is between the strands β4 and β5. The acetyl group is buried in the cleft and points toward the main chain amine of a residue on the strand β4. The structures shown in this figure include aminoglycoside 6'-N-acetyltransferases from *E. coli* (PDB ID: 2VQY-orange) and *E. faecium* (PDB ID: 1B87-yellow), a glyphosate N-acetyltransferase from *B. licheniformis* (PDB ID: 2JDD-pink), a serotonin N-acetyltransferase from sheep (PDB ID: 1CJW-green) and a histone N-acetyltransferase from *S. cerevisiae* (PDB ID: 1BOB-cyan). This figure was produced using Pymol (Schrodinger, 2010).

To compare the interactions made by acetyl coenzyme A with GNAT family members, the structures of the binary/ternary complexes of acetyl coenzyme A (or coenzyme A) with an aminoglycoside 6'-N-acetyltransferase from *E.coli* (PDB ID: 2VQY) and *E. faecium* (PDB ID: 1B87), a glyphosate N-acetyltransferase from *B. licheniformis* (PDB ID: 2JDD), a serotonin N-acetyltransferase from sheep (PDB ID: 1CJW) and a histone N-acetyltransferase from *S. cerevisiae* (PDB ID: 1BOB) were compared. From an analysis of these structures, it can be seen that the conformation of the acetyl coenzyme A in the binary complexes with various GNAT members is very similar. From the acetyl group to the pyrophosphate group the conformation of acetyl coenzyme A is generally uniform, with the pantotheine arm lying in the cleft between the strands β4 and β5. The carbonyl and amide groups of the pantotheine arm make hydrogen bonds to both of these strands, somewhat completing the hydrogen bonding of the central beta sheet. As BPSL0606 has a very similar structure to the other family members in the region of beta strands 4 and 5, these main chain interactions between the protein and the acetyl coenzyme A could still be made in a binary complex, despite differences in the amino acid sequence.

In contrast, there are some differences between various GNAT complex structures at the 3' phosphate ADP of the acetyl coenzyme A. In particular, in the structure of histone N-acetyltransferase from *S. cerevisae* (PDB ID: 1BOB), the adenine ring is flipped to the other side, compared to other structures (Figure 7.15).

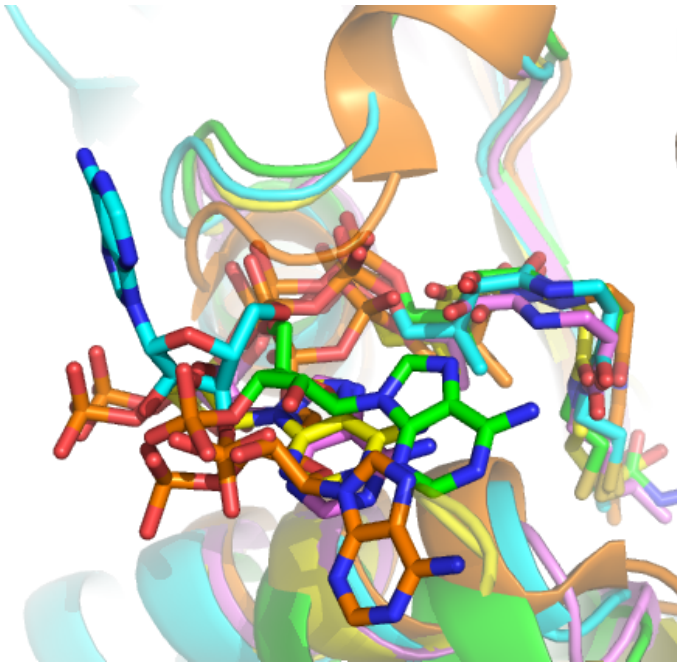


Figure 7.15 Different conformations of bound acetyl coenzyme A: The position of 3' phosphate ADP of the acetyl coenzyme A bound to GNAT structures has some variation. These structures shown in this figure include aminoglycoside 6'-N-acetyltransferases from *E. coli* (PDB ID: 2VQY-orange) and *E. faecium* (PDB ID: 1B87-yellow), a glyphosate N-acetyltransferase from *B. licheniformis* (PDB ID: 2JDD-pink), a serotonin N-acetyltransferase from sheep (PDB ID: 1CJW-green) and a histone N-acetyltransferase from *S. cerevisiae* (PDB ID: 1BOB-cyan). This figure was produced using Pymol (Schrodinger, 2010).

The majority of the contacts made between the family members and the pyrophosphate of acetyl coenzyme A are again hydrogen bonds to main chain amides. In addition, some members have interactions to side chains, with for example the side chains of Ser 88 and Thr 77 in a glyphosate-N-acetyltransferase from *B. licheniformis* (PDB ID: 2JDD), making hydrogen bonds to the pyrophosphate oxygens. In addition, the side chain of Arg 82 in an aminoglycoside 6'-N-acetyltransferase from *S. enteritidis* (PDB ID: 1S3Z) is close to the pyrophosphate, making stabilizing electrostatic interactions between the positively charged side chain and the negatively charged pyrophosphate. Although in many GNAT proteins the residue at an equivalent position to Thr 77

in the glyphosate-N-acetyltransferase (PDB ID: 2JDD) is a hydrophobic residue (Val or Leu) this residue is serine in the acyl coA N-acetyltransferase from *P. aeruginosa* (PDB ID: 1XEB) (Ser 83) and is also a threonine in BPSL0606 (Thr 83), perhaps indicating that equivalent interactions are made in these two proteins. At the Ser 88 position in the glyphosate N-acetyltransferase (PDB ID: 2JDD) there is much more variation in sequence, with Arg, Pro, His Lys, Thr and Gly all being present at this position, indicating that this residue is not essential for acetyl coenzyme A binding. The residue at this position in BPSL0606 is an aspartic acid (Asp 94) which at first glance may preclude the close association the negatively charged pyrophosphate to this acidic side chain. However, the side chain of this residue adopts many different conformations in the wider family, so may well point out into the solvent in a binary complex of acetyl coenzyme A with BPSL0606. In addition, Arg 121 in BPSL0606 (on the opposite side of the pyrophosphate from Asp 94, is in an ideal position to both coordinate and stabilise the negatively charged pyrophosphate.

The adenine ring of acetyl coenzyme A lies on the periphery of the protein, packing against the protein surface and can adopt a variety of different positions. The 3' phosphate attached to the ribose ring, is also in different positions in the various family member structures, however, there is usually one or more positively charged side chains in the vicinity, to stabilise this group. For example, Lys 120 and Lys 85 in the glyphosate N-acetyltransferase (PDB ID: 2JDD), Arg82 in a ribosomal S18 protein RimI in *S. typhimurium* (PDB ID: 2CNM) and Lys94 in the aminoglycoside 6'-N-acetyltransferases from *S. enteritidis* (PDB ID: 1S3Z). The equivalent residues to Lys 120 and Lys 85 in 2JDD are Arg 127 and Arg 91 in BPSL0606, respectively and could thus stabilise the 3' phosphate in a BPSL0606-acetyl coenzyme A complex.

In the light of the structural and functional similarity between BPSL0606 and the GNAT family members in the acetyl coenzyme A binding site, an acetyl coenzyme A molecule was modeled into the active site cleft of BPSL0606 (Figure 7.16). As can be seen, the coenzyme fits the structure well. It must be noted, however, that the relative position of helix α4 is slightly different in BPSL0606 compared to the other family members and perhaps may adjust its position on

coenzyme binding as the case in serotonin N-acetyltransferase structures. The crystal structure of the bisubstrate analogue bound serotonin N-acetyltransferase from sheep (Figure 7.17b) revealed a conformational change in the protein, compared to the apo enzyme (Figure 7.17a). When the acetyl coenzyme A binding site is occupied, the protein is reordered to allow the ligand access. In this case, the α-helix α1 was dramatically extended in length, substituting the short β-strand and shortening the loop connected to the α-helix, α2, which was also extended (Figure 7.17c).

To model an acetyl coenzyme A into BPSL0606, the structure of aminoglycoside 6'-N-acetyltransferase (AAC6') from *E. faecium* (PDB ID: 1B87) in complex with acetyl coenzyme A was used to compare the active site as the acetylation of aminoglycoside contributes to antibiotic resistance in bacteria which may be the case in BPSL0606.

The main chain interactions made between acetyl coenzyme A and AAC6', include the amines of five residues in the protein, which are H-bonded to the phosphates and pantothenic acid in the acetyl coenzyme A (Table 7.5). The α-phosphate oxygen atoms in acetyl coenzyme A are H-bonded to the amines of Gln 86 and Thr 89 and the β-phosphate oxygens are H-bonded to the amines of Lys 84 and Gly 88. Also the oxygen atom in pantothenic acid interacts with the amine of Val 78, which is equivalent to Thr 83 in BPSL0606. Since the backbone of this region is similar to BPSL0606 with the exception that the peptides are flipped for the residues Lys 84 and Gln 86, it is possible that acetyl coenzyme A could interact with BPSL0606 in similar manner. Unlike other acetyl coenzyme A bound GNATs structures, this complex lacks a H-bond between the acetyl group and the main chain (equivalent to residue Phe 220 in histone N-acetyltransferase, PDB ID: 1BOB) due to the low positional accuracy of the acetyl group.

Table 7.5 The main chain interaction of the acetyl coenzyme A bound to AAC6'

Acetyl coenzyme A	AAC6' (PDB ID: 1B87)	BPSL0606
α-phosphate	Gln 86 (NH)	Arg 91 (NH)
	Thr 89 (NH)	Asp 94 (NH)
β-phosphate	Lys 84 (NH)	Asp 89 (NH)
	Gly 88 (NH)	Ile 93 (NH)
Pantotheine	Val 78 (NH)	Thr 83 (NH)

Side chain interactions with acetyl coenzyme A, include both H-bonds made with the phosphate groups and hydrophobic contacts with the pantotheine arm and the acetyl portion of the acetyl coenzyme A molecule. Side chain residues of the AAC6' involved in the acetyl coenzyme A contacts are listed in Table 7.6. The Lys 149 side chain in ACC6' interacts with the 3'phosphate, this residue is equivalent to Arg127 in BPSL0606, which could make a H-bond as can be seen in other GNATs. The hydroxyl group of Thr 89 in AAC6' makes a contact to the α-phosphate of acetyl coenzyme A, which could also be possible using the equivalent residue, Asp 94, in BPSL0606. The pantothenic acid and β-mercaptoethylamine moieties in the acetyl coenzyme A are directed into a hydrophobic pocket made from the side chains of Phe 146, Leu 76, Val 77 in AAC6'. The side chain of Tyr 147 can also make an interaction with the thiol group of acetyl coenzyme A. Additionally the hydrophobic side chains of Leu 73 and Leu 109 develop the binding pocket deep enough to accommodate the acetyl moiety of the acetyl coenzyme A. In the V-cleft of BPSL0606, the hydrophobic pocket could also be used to bind acetyl coenzyme A, as the residues involved are similar to those in the AAC6'.

These AAC6' residues involved in these interactions and the main chain conformation in the acetyl coenzyme A binding site are conserved in BPSL0606, indicating that BPSL0606 could possibly bind acetyl coenzyme A in similar manner (Figure 7.16).

Table 7.6 Comparison of residues involved in side chain interaction made between AAC6' and acetyl coenzyme A and equivalent residues in BPSL0606

Interactions with AcCoA	Residues involved	
	AAC6' (PDB ID 1B87)	BPSL0606 equivalent residue
3'-phosphate	Lys 149	Arg 127
α-phosphate	Thr 89	Asp 94
hydrophobic pocket for pantotheine	Phe 146	Phe 124
	Leu 76	Ile 81
	Val 77	Leu 82
	Tyr 147	Tyr 125
hydrophobic pocket for the acetyl	Tyr 147	Tyr 125
	Leu 73	Ile 78
	Leu 109	Val 115

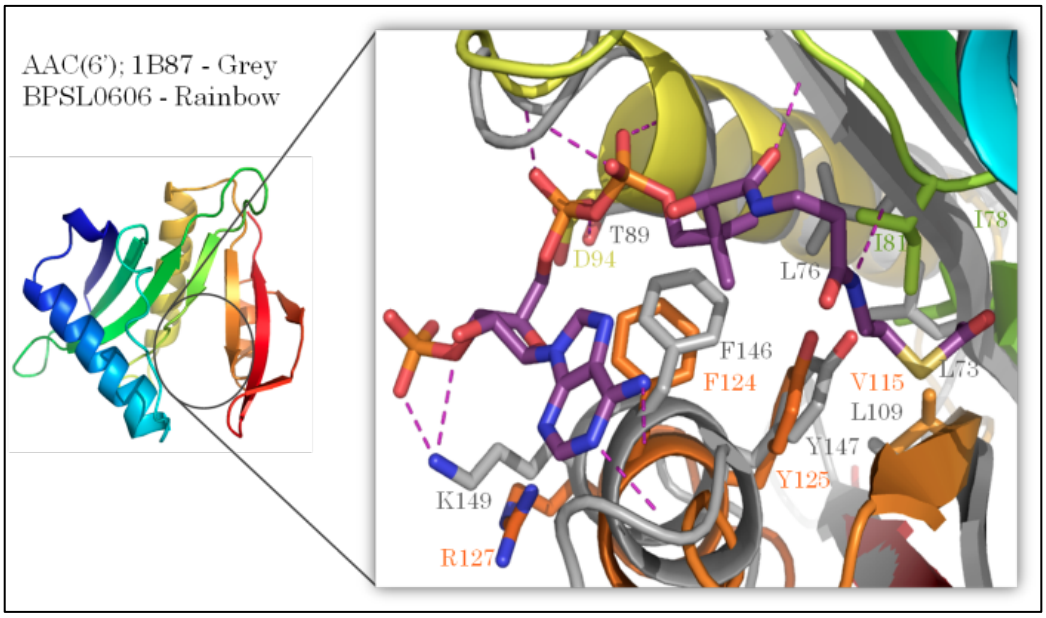
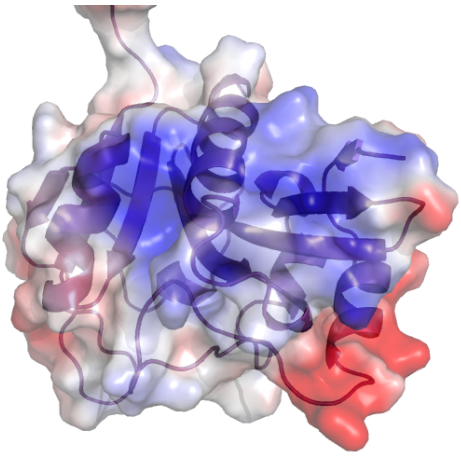
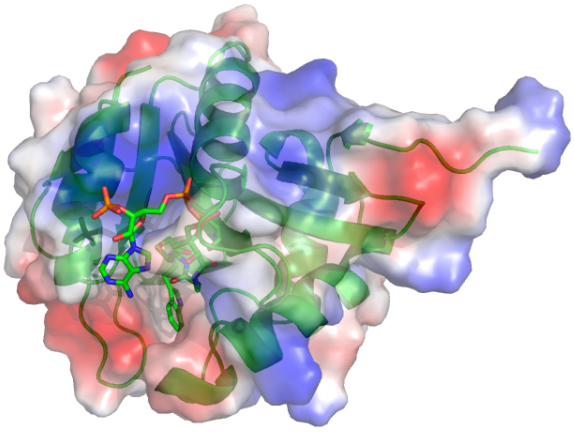


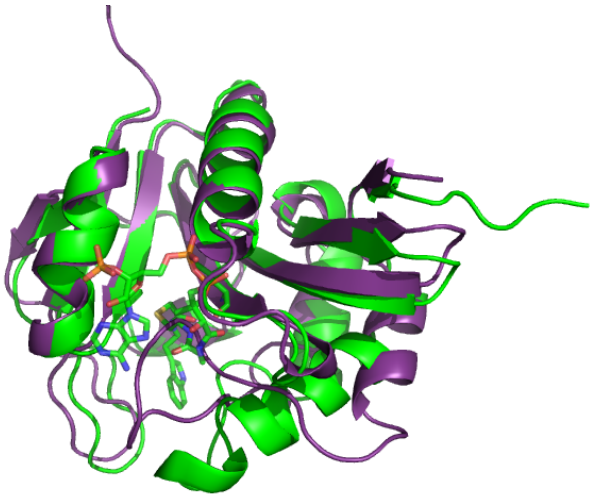
Figure 7.16 Superposition of the BPSL0606 and the aminoglycoside 6' N-acetyltransferase from *E. faecium* (PDB ID: 1B87): The structure of BPSL0606 (rainbow color) is superimposed on an aminoglycoside 6'-N-acetyltransferase (grey) with acetyl coenzyme A (purple) with a Cα-RMSD 1.3Å, showing the residues involving in the acetyl coenzyme A binding. This figure was produced using Pymol (Schrodinger, 2010).



a)



b)



c)

Figure 7.17 The conformational change of serotonin N-acetyltransferase: Surface electrostatics of a) unbound structure (PDB ID: 1b6b, violet) and b) bisubstrate analogue bound structure (PDB ID: 1b87, green), showing the different conformation of serotonine N-acetyltransferase when the acetyl coenzyme A binding site is occupied. c) Superposition of both structures shows the different secondary structure. This figure was produced using Pymol (Schrodinger, 2010).

7.7. Analysis of substrate binding of GNATs

In the N-acetylation reaction catalysed by the GNAT family, there is a direct nucleophilic attack of the amino group of the acceptor on the thioacyl carbon, leading to breakage of the S-C bond (Dyda *et al*, 2000). For this to occur, the amino group must be deprotonated, and thus there is often a general base close to the acetyl group to facilliate this. For example, in *M.tuberculosis* Rv1347c, His 130 and Asp 168 lie on opposite sides of the cleft, close to the acetyl group (Card *et al*, 2005). In BPSL0606 these residues are Asp 79 and Gln 116, respectively, and the side chain of Asp 79 would be ideally positioned to act as a general base to abstract a proton from the acceptor amino group.

Since acceptor substrates of the GNAT superfamily across species are widely diverse, it is difficult to predict the substrate of BPSL0606 from the structure alone. However, one of the functions of GNAT members from pathogenic bacteria, is to acetylate aminoglycoside antibiotics, as part of an antbiotic resistance strategy. A number of enzymes in this class were selected to analyse the binding of aminoglycoside substrates in order to try to model the substrate-binding site in BPSL0606. These include crystal structures of aminoglycoside N-acetyltransferases bound to aminoglycosides in the Gram-negative bacteria.

Aminoglycoside N-acetyltransferases play an important catalytic role in the regioselective acetylation of one amine group in amionoglycoside antibiotics, which disrupt the bacterial protein synthesis by binding to the A-site of 16S rRNA (Vetting *et al*, 2008b). This enzymatic modification by N-acetylation of the active aminoglycosides inhibits antibiotic binding, resulting in difficulties in infectious disease treatment regimes. Aminoglycoside 6' N-acetyltransferase (AAC6') is the most common aminoglycoside-modifying enzyme found in most Gram-negative bacteria (Vakulenko & Mobashery, 2003). Crystal structures of AAC6' bound to acetyl coenzyme A or coenzyme A and aminoglycoside substrates in *E. coli* (Vetting *et al*, 2008b) were chosen as models in an analysis of substrate binding.

Crystal structures of ternary complexes of AAC6' from *E. coli* with acetyl coenzyme A or coenzyme A and different aminoglycosides; kanamycin (PDB ID: 1V0C), parmomycin (PDB ID: 2VQY) and ribostamycin (PDB ID: 2BUE),

revealed that the aminoglycoside substrate binding site of AAC6' is in a cavity made by the anti-parallel β-sheet; β2, β3, β4 with α-helices; α1, α2 and the loop connected between β6 and β7 (Figure 7.18). In aminoglycosides, the 6-amino hexose ring, which is the target of 6' N-acetylation, was buried in the negatively charge active site with the 6' amine group pointing to the acetyl group of the acetyl coenzyme A. Asp 115 and Asp 152 are the crucial residues that interact with 3' amine group of the 2-deoxystreptamine ring and 6' amine group of the amino sugar, determining the substrate specificity (Figure 7.19).

The ternary complex of AAC6' from *E. coli* with acetyl coenzyme A and kanamycin was used to model the substrates into the BPSL0606 structure as both these structures share similar surface electrostatics in the active site. (Figure 7.20) In the complex structure, interactions are made between the main chain carbonyl of Ser 98 in the AAC6' with 3"NH_3 of kanamycin and the side chain of Asp 100 with 3"NH_3 and 4"-OH. This loop region is not displaced in BPSL0606, therefore, it is not possible to have these interactions in BPSL0606. However, the crucial residues for the binding of aminoglycoside, described above, are present in the same position in BPSL0606. The residues involved in substrate binding in AAC6' (Asp 115 and Asp 152) are similar in BPSL0606 which are Asp 79 and Gln 116, respectively, and the side chains of these two residues could be H-bonded with the 3' amine group of the 2-deoxystreptamine ring and 6' amine group of the amino sugar in the kanamycin, similar to the AAC6' complex structure (Figure 7.21).

The substrates acetyl coenzyme A and kanamycin were modeled into the BPSL0606 structure to propose a putative active site for this protein (Figure 7.22 and 7.23). Although both acetyl coenzyme A and kanamycin fit in the active site groove of BPSL0606, it is by no means certain that these are the actual substrates of this protein and further crystallographic and biochemical analyses must be carried out to determine the function of BPSL0606.

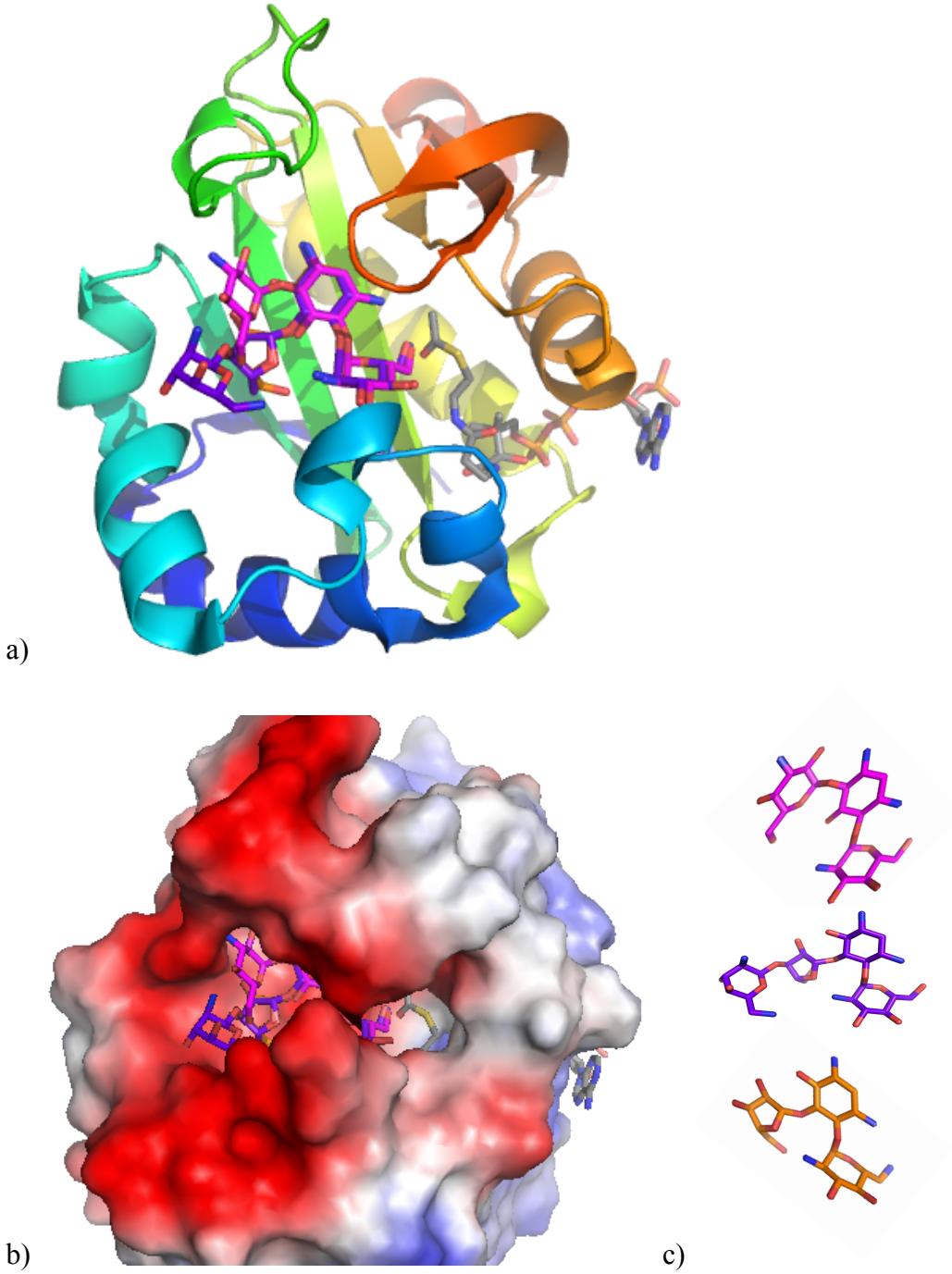


Figure 7.18 Crystal structure of the aminoglycoside 6' N-acetyltransferase from *E. coli* complexed with acetyl coenzyme A and different aminoglycoside substrates: a) A cartoon representation of the structure of AAC6' colored from N-terminus (blue) to C-terminus (red) are displayed with bound substrates. b) Surface electrostatics diagram of AAC6' (negative charge in red and positive charge) shows the active sites within the cavity where the substrates were bound to the enzyme. c) Aminoglycoside substrates; kanamycin (pink), parmomycin (purple), ribostamycin (orange). This figure was produced using Pymol (Schrodinger, 2010).

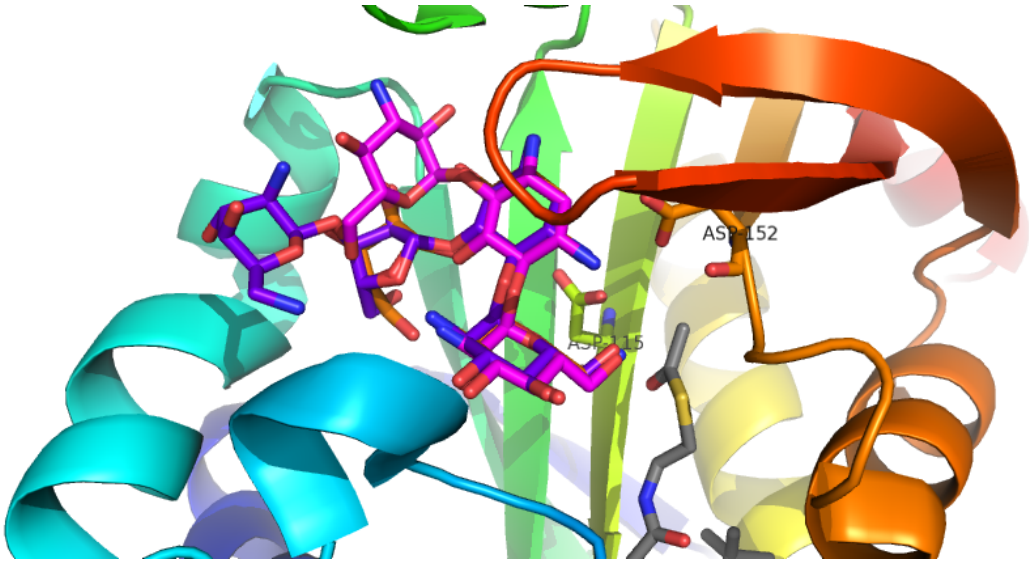
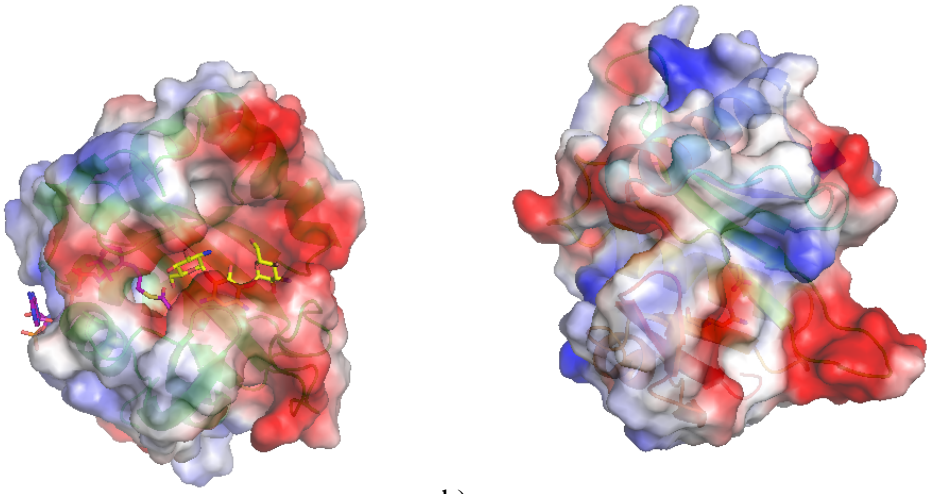


Figure 7.19 The active site of AAC6': Asp 115 and Asp 152 in the AAC6' are significant determinants for aminoglycoside substrate specificity in the catalysis of 6'N-acetylation. This figure was produced using Pymol (Schrodinger, 2010).



a) b)

Figure 7.20 Surface electrostatics of the structures of AAC6' complex with substrates (a) and BPSL0606 (b): An opening tunnel in BPSL0606 could possible be occupied by the acetyl coenzyme A and kanamycin, similar to the AAC6'. This figure was produced using Pymol (Schrodinger, 2010).

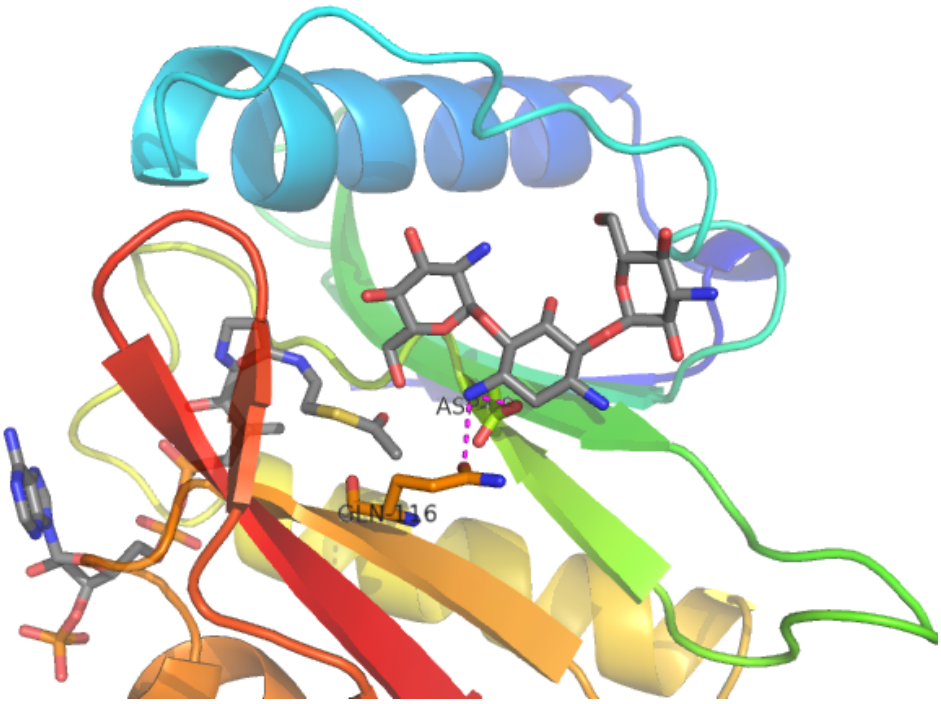


Figure 7.21 The proposed kanamycin binding site of BPSL0606: Gln 116 and Asp 79 in BPSL0606 could stabilise a substrate such as kanamycin in the active site. This figure was produced using Pymol (Schrodinger, 2010).

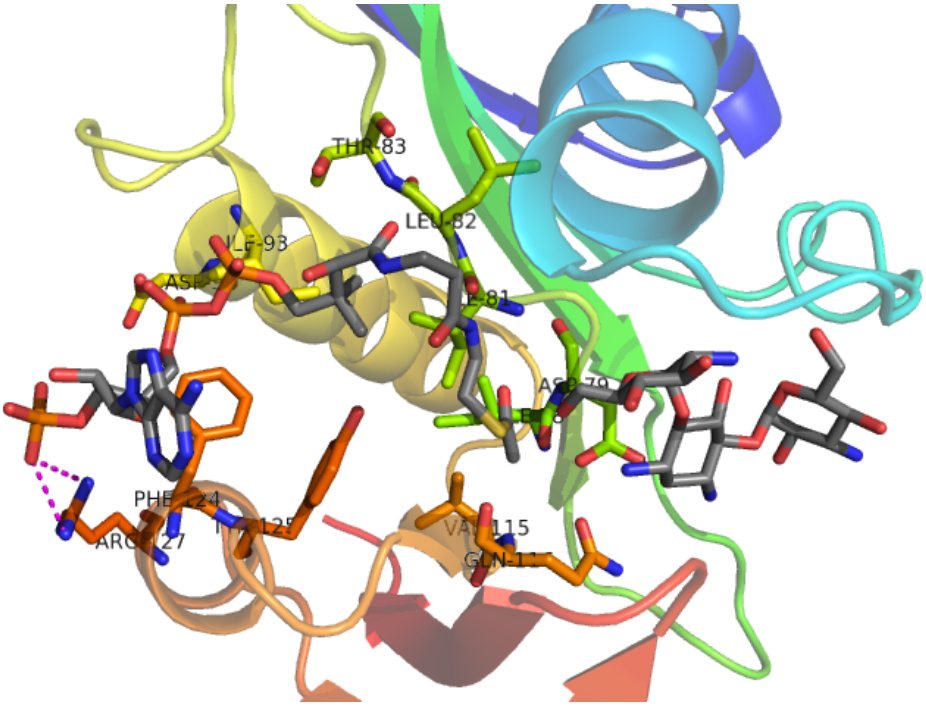


Figure 7.22 The proposed acetyl coenzyme A binding site of BPSL0606: Arg 127 is H-bonded to the 3'phosphate of the acetyl coenzyme A, and the pantotheine and acetyl group lie in the hydrophobic pocket. This figure was produced using Pymol (Schrodinger, 2010).

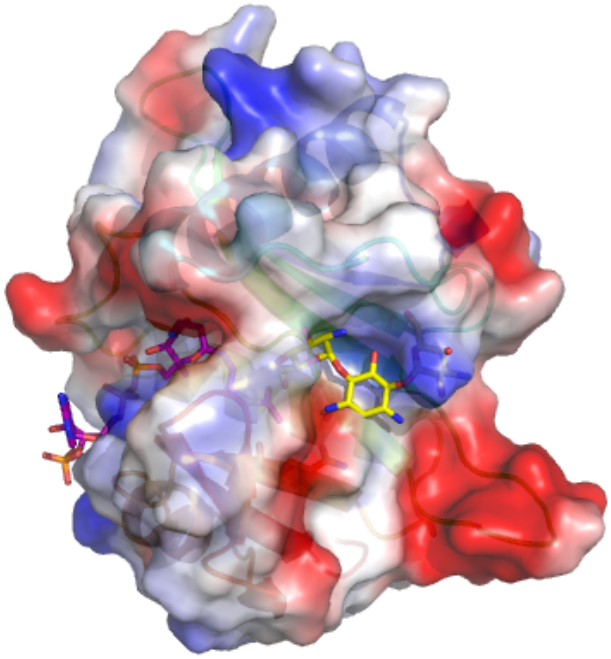


Figure 7.23 Model of BPSL0606 complex with acetyl coenzyme A and kanamycin: Both acetyl coenzyme A and kanamycin could occupy the cavity, which may well be the active site of this protein. This figure was produced using Pymol (Schrodinger, 2010).

7.8. Chapter summary

Analysis of the BPSL0606 structure has shown that it is structurally similar to the GCN-5 related N-acetyltransferase superfamily, in which these proteins use acetyl coenzyme A as a functional group donor. Known structures of these superfamily members with acetyl coenzyme A were examined and superposed with the BPSL0606 structure. Conserved residues among the GNAT family were identified and the equivalent residues in BPSL0606 were shown to be similar. Cocrystallisation of BPSL0606 and acetyl coenzyme A were set up, but X-ray data sets showed that acetyl coenzyme A did not appear in these crystals. Also, ligand observation by 1D ^{1}H NMR suggested that BPSL0606 did not bind to acetyl coenzyme A. Despite this, the proposed complex of BPSL0606, acetyl coenzyme A and kanamycin can be modeled at the putative binding site in BPSL0606.

Chapter 8: Conclusions and Future work

This chapter summarises the work that contributed to this PhD research. From the results obtained, the findings of this thesis provide some informative data as part of a directed structural genomics project of *B. pseudomallei*. Suggestions for what could be further investigated are also included.

8.1. Summary

Since little is known about molecular mechanisms and virulence in *B. pseudomallei*, the potential targets from previously identified putative essential genes and immunogenic proteins were selected for this PhD research in order to gain more knowledge and informative data from structural studies. Beginning with seven target genes, four full-length genes (BPSS0945, BPSS0603, BPSS0683, BPSL0606) and one truncated gene (BPSL3022) were successfully cloned. Two of them, BPSS0945 and BPSL0606, were expressed in a soluble form, which enabled protein crystallisation trials. BPSS0945, a putative subfamily M23 unassigned peptidase was unstable, which resulted in a non-homogenous protein sample and thus caused crystallisation failure. Only one protein, BPSL0606, crystallised sucessfully which enabled a structure to be determined by protein crystallography.

The success rate in each step is illustrated in Figure 8.1. This 14% success rate compared fairly favourable with a predicted rate of 10-20% for a straightforward a bacterial protein crystal structure determination (Rupp, 2010). Success rates in each step of structure determination from large structural genomic consortia over the world are ranging from one-third to two-thirds of the time (Terwilliger *et al*, 2009). According to published data from worldwide structural genomic centers, crystal structures have been obtained approximately 3% of total cloned targets (http://targetdb-dev.rutgers.edu/statistics/TargetStatistics.html). Soluble protein expression and crystallisation are still challenging tasks to feed crystallographic work, despite the fact that 15-20% of less than 50 kDa non membrane proteins are likely to be suitable for structural biology (Edwards *et al*, 2000).

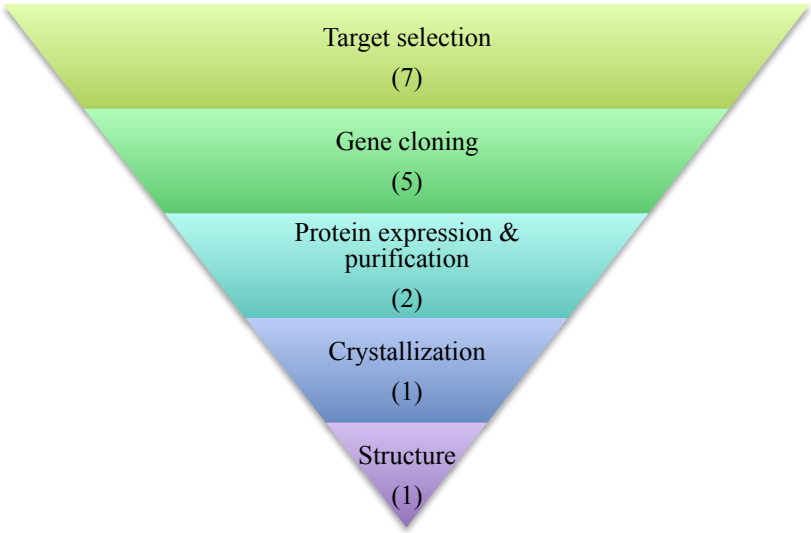


Figure 8.1 Pyramid of attrition rate in key steps of the structure determination project: From target selection, the molecular cloning has a reasonable success rate but soluble protein expression and crystallisation have a major impact on crystallographic studies. Numbers of genes/proteins in this project that were successful in each step are shown in parentheses.

BPSL0606 was successfully cloned, expressed as a full-length protein and purified. Analysis of the purified protein by mass spectrometry confirmed the presence of full-length protein (23.5 kDa), prior to crystallisation trials. Native and selenomethionine substituted protein crystallised in space group $P2_1$ from solutions of 14-24% PEG3350, 0.1 M Bis-tris propane pH 6.3-6.5 and 0.2 M potassium thiocyanate, yielding reasonable size single long rod crystals suitable for data collection. The crystal structure of BPSL0606 was determined using MAD method, to a resolution of 2.2 Å. The 65 C-terminal residues of the protein were not visible in the electron density map and the crystals were investigated to see if any proteolysis had occurred. SDS PAGE of dissolved crystals showed that the molecular weight had reduced to 17 kDa, indicating some truncation of the protein, by an, as yet unknown, mechanism.

Attempts were made to crystallise the truncated protein, from a protein sample that was two months old. The mass spectrometry analysis of this sample showed it to have a molecular weight of 17 kDa consistent with it representing residues 1-159. This sample did not crystallise in the same way as the original "full-length" sample, forming crystals both under different conditions (25% PEG3350, 0.1 M Bis-tris pH 5.5 and 0.2 M ammonium sulphate) and also in a different space group ($P2_12_12_1$, a=, 60.9 b=75, c=78.2 Å, α=β=γ=90°). The crystal structure obtained from the truncated protein was determined at 2.2 Å resolution by molecular replacement using the previous structure.

These two structures of BPSL0606 had a similar conformation but showed differences in crystal packing in the different layers. The BPSL0606 $P2_1$ crystal form had four copies in the asymmetric unit, whereas the $P2_12_12_1$ crystal form had two molecules in the asymmetric unit.

The crystal structure of BPSL0606 consists of eight β-strands and three α-helices, including a turn of 3_{10}-helix. Structure comparison to the published known structures in the protein data bank revealed that BPSL0606 is structurally similar to the GCN5-related N-acetyltransferase (GNAT) superfamily. Structure based alignment of BPSL0606 the known GNATs structures showed the conservation of residues in four signature motifs of the GNAT superfamily.

The dimer formation in crystal packing agrees with gel filtration analysis, which predicted that BPSL0606 is likely to be a dimeric protein, not a monomer. The dimer interface in the BPSL0606 structure buries approximately 11% of the subunit surface area. Two of the residues of the dimer interface (Leu 11 and Ser 12) are also conserved among the other similar GNAT dimers. The BPSL0606 interface is also similar to a number of interfaces that belong to GNAT superfamily members.

From structure analysis, binding of acetyl coenzyme A seemed possible for BPSL0606, as residues involved in the acetyl coenzyme A binding in GNATs are also conserved in BPSL0606. The V-cleft made by strands β4 and β5 in BPSL0606 also provides possible tunnel that could accommodate acetyl

coenzyme A as is seen in the GNAT family members. However, co-crystallisation of BPSL0606 and acetyl coenzyme A suggested that BPSL0606 did not bind to the acetyl coenzyme A and there was also no indication of binding by acetyl coenzyme A and BPSL0606 from ligand observation by H^1NMR. A key residue in the acetyl coenzyme A binding in the GNAT family is a glycine or alanine lying in the pocket. The equivalent residue in BPSL0606 is an isoleucine, Ile 93, and this large side chain may prevent acetyl coenzyme A binding. However, it must also be mentioned that binary complexes cannot always be crystallised for other GNAT members, for example, Rv1347c, a putative antibiotic resistance protein from *M. tuberculosis*.

Since BPSL0606 may be related to one of the GNAT members, aminoglycoside N-acetyltransferases, which are involved in antibiotic resistant mechanisms, attempts were made to model acetyl coenzyme A and kanamycin into the structure of BPSL0606 and a putative active site in BPSL0606 was proposed. As the crucial residues for aminoglycoside binding are conserved in BPSL0606 (Asp 79 and Gln 116), it is possible that kanamycin could accommodate the putative binding site even though there is much less negative charge than the aminoglycoside N-acetyltransferase.

8.2. Future work

Although the findings in this thesis, provided useful molecular and structural information for *B. pseudomallei*, there are still fundamental research questions awaiting further investigations.

As it has proved difficult to show that BPSL0606 binds to acetyl coenzyme A, further investigations could include a structure comparison to the acyl homoserine lactone (AHL) synthase (LasI) structure from *P. aeruginosa* (PDB ID: 1RO5). This structure is also similar to BPSL0606, however, LasI catalyses a different reaction from other members of the GNAT superfamily. Unlike GNAT members, LasI uses S-adenosyl-L-methionine (SAM) and acyl acyl-carrier protein (acyl ACP) as substrates (Gould *et al*, 2004). It is possible, therefore, that BPSL0606 also uses SAM and acyl ACP as binding partners, and cocrystallisation with

different substrates including SAM, acyl-ACP and their analogues would be carried out to further characterise the active site of BPSL0606. The ternary complex structure of TofI from *B. glumae* and substrates (Chung *et al*, 2011) could also be useful for analysis of potential substrate binding sites. Superposition of these structures on to the BPSL0606 structure would assist an identification of putative substrate binding sites in BPSL0606.

In addition, a ligand binding assay may also provide a clue in biological function of BPSL0606 and could lead to the rational drug and/or inhibitor design. A thermal shift assay can be employed to assist a screening of possible substrates that could be bound to BPSL0606 and identify the protein-ligand interaction.

A pull down assay of an affinity tagged BPSL0606 against *B. pseudomallei* cell extract would be done in order to determine whether cellular molecules, for example signaling molecules like AHL, could interact with BPSL0606. Due to the laboratory safety regulation of the harmful pathogen, BPSL0606 homologues in closely non-pathogenic strain such as *B. thailandensis* and *B. cenocepacia* can be substituted for the biochemical studies. For these studies it may be beneficial to clone the equivalent protein to BPSL0606 from the non-pathogenic strains of *Burkholderia*.

To clarify sequence differences seen in BPSL3022, BPSS0945, BPSS0603 and BPSL0606 whether these have occurred as a result of a mutation from PCR or a real strain variation, the amplification and cloning of these genes should be repeated and a number of recombinant clones should be sequenced.

Further work on other target proteins studied in this thesis could involve:
-BPSL3022 cell division protein FtsQ: Making different constructs and optimizing the protein expression condition to obtain the soluble protein for crystallisation.

-BPSS0945 subfamily M23B unassigned peptidases: Making different constructs without the possible signal peptide region to produce the mature protein which may be more stable than the full-length protein. In addition, to inactivate the autolysis of the protein, protease inhibitors could be used.

-BPSS0603: A new clone without sequence difference could be made and the soluble protein expression could possibly be observed.

-BPSS0683: Soluble protein expression could be achieved by co-expression of molecular chaperones such as GroES-GroEL, DnaK-DnaJ-GrpE and ClpB, which could assist the *in vivo* protein folding.

Appendix

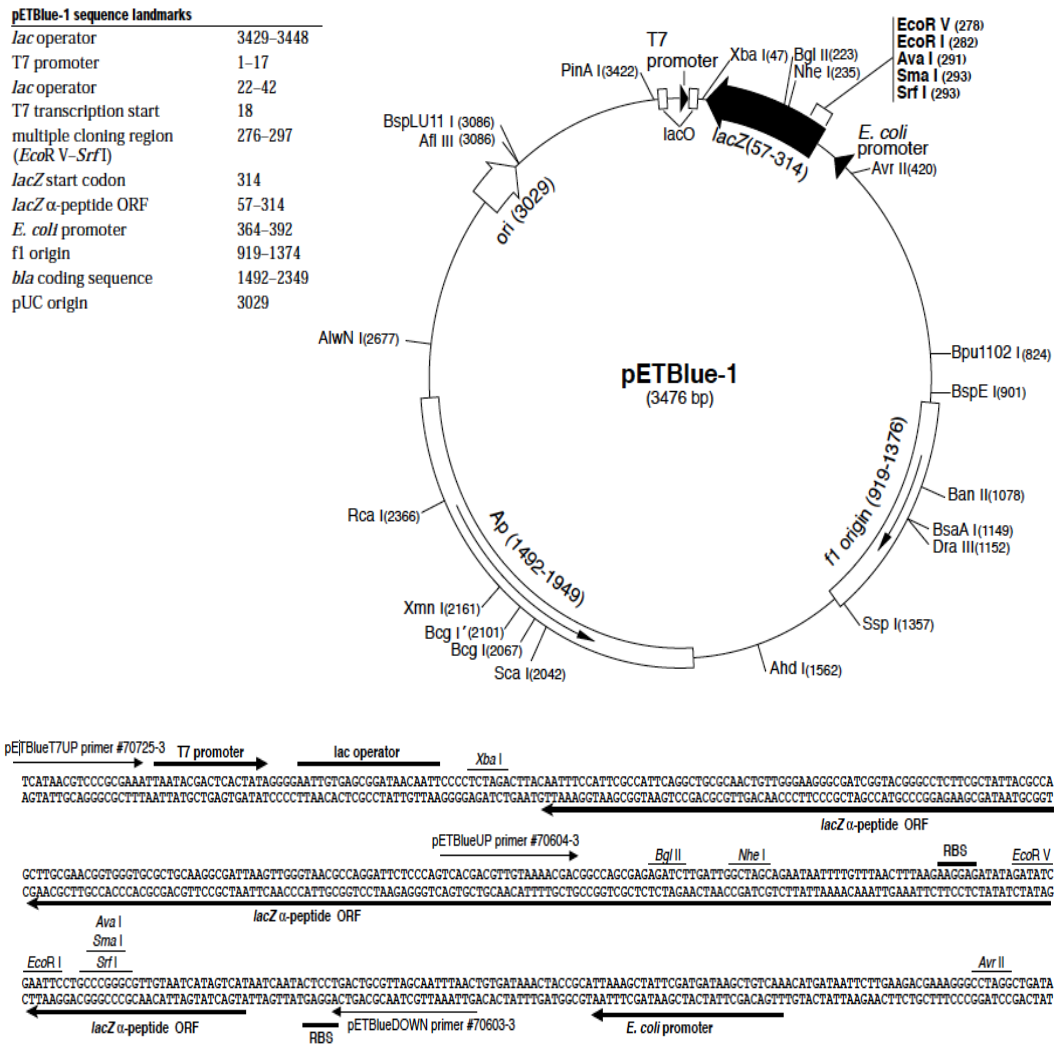
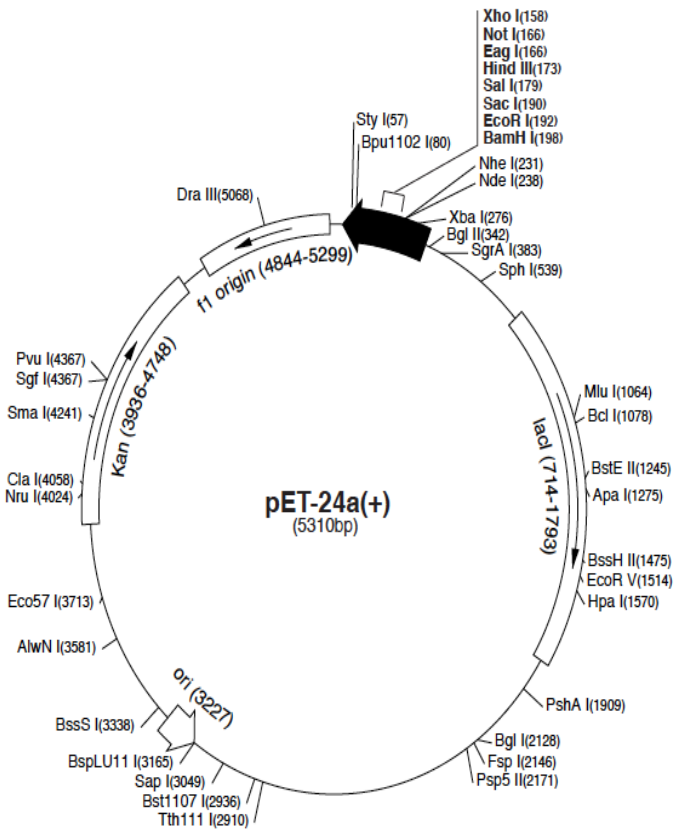


Figure 3.1 The pETBlue-1 vector map and cloning and expression region. The vector contains a pre-digested *Eco*RV cloning site which provides optimal space from the ribosome bind site (RBS) for efficient translation initiation when the ligated insert begins with 5' ATG. The protein expression is driven by a strong T7 RNA polymerase promoter with an inducible *lac* apparatus.

pET-24a(+) sequence landmarks	
T7 promoter	311-327
T7 transcription start	310
T7•Tag coding sequence	207-239
Multiple cloning sites	
(*Bam*H I - *Xho* I)	158-203
His•Tag coding sequence	140-157
T7 terminator	26-72
lacI coding sequence	714-1793
pBR322 origin	3227
Kan coding sequence	3936-4748
f1 origin	4844-5299

The maps for pET-24b(+), pET-24c(+) and pET-24d(+) are the same as pET-24a(+) (shown) with the following exceptions: pET-24b(+) is a 5309bp plasmid; subtract 1bp from each site beyond *Bam*H I at 198. pET-24c(+) is a 5308bp plasmid; subtract 2bp from each site beyond *Bam*H I at 198. pET-24d(+) is a 5307bp plasmid; the *Bam*H I site is in the same reading frame as in pET-24c(+). An *Nco* I site is substituted for the *Nde* I site with a net 1bp deletion at position 238 of pET-24c(+). As a result, *Nco* I cuts pET24d(+) at 234, and *Nhe* I cuts at 229. For the rest of the sites, subtract 3bp from each site beyond position 239 in pET-24a(+). *Nde* I does not cut pET-24d(+). Note also that *Sty* I is not unique in pET-24d(+).



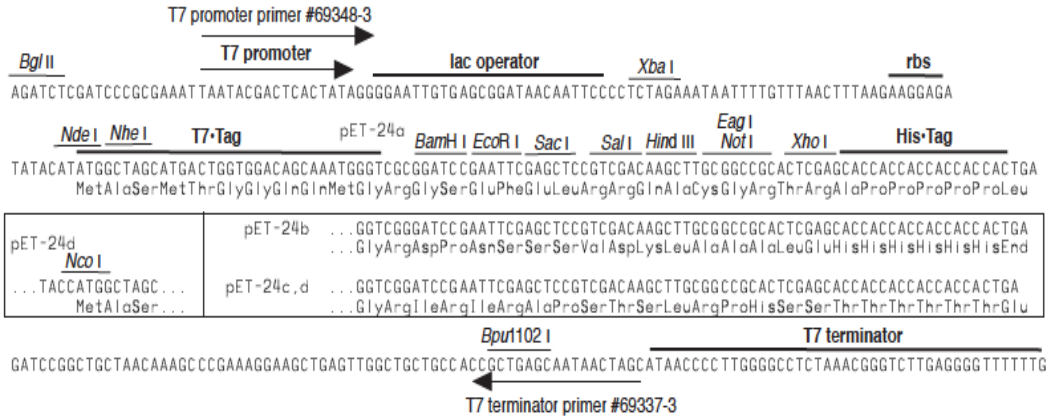


Figure 3.2 The pET24-a vector map and cloning and expression region. The pET24-a vector contains multiple cloning sites (RBS) with a kanamycin selective marker and also provides an optional C-terminal His tag fusion. It facilitates protein expression driven by a strong T7 RNA polymerase promoter with an inducible *lac* apparatus.

References

Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography* **66:** 213-221

Adler NRL, Govan B, Cullinane M, Harper M, Adler B, Boyce JD (2009) The molecular and cellular basis of pathogenesis in melioidosis: how does *Burkholderia pseudomallei* cause disease? *Fems Microbiol Rev* **33:** 1079-1099

Akira S, Uematsu S, Takeuchi O (2006) Pathogen recognition and innate immunity. *Cell* **124:** 783-801

Allwood EM, Devenish RJ, Prescott M, Adler B, Boyce JD (2011) Strategies for Intracellular Survival of *Burkholderia pseudomallei*. *Frontiers in microbiology* **2:** 170-170

Ashdown LR (1979) An improved screening technique for isolation of *Pseudomonas pseudomallei* from clinical specimens. *Pathology* **11:** 293-297

Bartpho T, Wongsurawat T, Wongratanacheewin S, Talaat AM, Karoonuthaisiri N, Sermswan RW (2010) The diversity of genomic islands among *Burkholderia pseudomallei* clinical and environmental isolates. In *VIth World Melioidosis Congress 2010*. Townsville, Australia

Bhatnagar RS, Futterer K, Waksman G, Gordon JI (1999) The structure of myristoyl-CoA:protein N-myristoyltransferase. *Biochim Biophys Acta* **1441:** 162-172

Blow DM (2002) *Outline of crystallography for biologists*, Oxford: Oxford : Oxford University Press, 2002.

Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* **72:** 248–254

Brannigan JA, Smith BA, Yu Z, Brzozowski AM, Hodgkinson MR, Maroof A, Price HP, Meier F, Leatherbarrow RJ, Tate EW, Smith DF, Wilkinson AJ (2010) N-myristoyltransferase from *Leishmania donovani*: structural and functional characterisation of a potential drug target for visceral leishmaniasis. *J Mol Biol* **396:** 985-999

Brenner SE (2001) A tour of structural genomics. *Nat Rev Genet* **2:** 801-809

Brunger AT (1992) Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355:** 472-475

Burk DL, Ghuman N, Wybenga-Groot LE, Berghuis AM (2003) X-ray structure of the AAC(6)-Ii antibiotic resistance enzyme at 1.8 Å resolution; examination of oligomeric arrangements in GNAT superfamily members. *Protein Sci* **12:** 426-437

Burtnick MN, Brett PJ, Harding SV, Ngugi SA, Ribot WJ, Chantratita N, Scorpio A, Milne TS, Dean RE, Fritz DL, Peacock SJ, Prior JL, Atkins TP, DeShazer D (2011) The Cluster 1 Type VI Secretion System Is a Major Virulence Determinant in *Burkholderia pseudomallei*. *Infection and Immunity* **79:** 1512-1525

Card GL, Peterson NA, Smith CA, Rupp B, Schick BM, Baker EN (2005) The Crystal Structure of Rv1347c, a Putative Antibiotic Resistance Protein from *Mycobacterium tuberculosis*, Reveals a GCN5-related Fold and Suggests an Alternative Function in Siderophore Biosynthesis. *J Biol Chem* **280:** 13978-13986

Carson MJ, Barondess J, Beckwith J (1991) The FtsQ protein of *Escherichia coli*: Membrane topology, abundance, and cell division phenotypes due to overexpression and insertion mutations. *J Bacteriol* **173:** 2187-2195

Chan YY, Bian HS, Tan TMC, Mattmann ME, Geske GD, Igarashi J, Hatano T, Suga H, Blackwell HE, Chua KL (2007) Control of quorum sensing by a *Burkholderia pseudomallei* multidrug efflux pump. *Journal of Bacteriology* **189:** 4320-4324

Chan YY, Tan TMC, Ong YM, Chua KL (2004) BpeAB-OprB, a Multidrug Efflux Pump in *Burkholderia pseudomallei*. *Antimicrob Agents Ch* **48:** 1128–1135

Chantratita N, Sim B, Wuthiekanun V, Limmathurotsakul D, Amornchai P, Thanwisai A, Holden MT, Day NP, Tan P, Peacock SJ (2010). Development of ceftazidime resistance by *Burkholderia pseudomallei* following human infection associated with large-scale genomic deletion. *VIth World Melioidosis Congress 2010*; Townsville, Australia.

Chantratita N, Wuthiekanun V, Boonbumrung K, Tiyawisutsri R, Vesaratchavest M, Limmathurotsakul D, Chierakul W, Wongratanacheewin S, Pukritiyakamee S, White NJ, Day NPJ, Peacock SJ (2007a) Biological relevance of colony morphology and phenotypic switching by *Burkholderia pseudomallei*. *Journal of Bacteriology* **189:** 807-817

Chantratita N, Wuthiekanun V, Thanwisai A, Limmathurotsakul D, Cheng AC, Chierakul W, Day NPJ, Peacock SJ (2007b) Accuracy of enzyme-linked immunosorbent assay using crude and purified antigens for serodiagnosis of melioidosis. *Clin Vaccine Immunol* **14:** 110-113

Chayen NE, Saridakis E (2008) Protein crystallization: from purified protein to diffraction-quality crystal. *Nat Methods* **5:** 147-153

Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography* **66:** 12-21

Cheng AC, Currie BJ (2005) Melioidosis: Epidemiology, pathophysiology, and management. *Clinical Microbiology Reviews* **18:** 383-416

Cheng AC, Wuthiekanun V, Limmathurosakul D, Wongsuvan G, Day NPJ, Peacock SJ (2006) Role of selective and nonselective media for isolation of *Burkholderia pseudomallei* from throat swabs of patients with melioidosis. *J Clin Microbiol* **44:** 2316-2316

Chong CE, Lim BS, Nathan S, Mohamed R (2006) In silico analysis of *Burkholderia pseudomallei* genome sequence for potential drug targets. *In Silico Biol* **6:** 341-346

Chua KL, Chan YY, Gan YH (2003) Flagella are virulence determinants of *Burkholderia pseudomallei*. *Infection and Immunity* **71:** 1622-1629

Chuaygud T, Tungpradabkul S, Sirisinha S, Chua KL, Utaisincharoen P (2008) A role of *Burkholderia pseudomallei* flagella as a virulent factor. *T Roy Soc Trop Med H* **102 Suppl 1:** S140-144

Chung J, Goo E, Yu S, Choi O, Lee J, Kim J, Kim H, Igarashi J, Suga H, Moon J, Hwang I, Rhee S (2011) Small-molecule inhibitor binding to an N-acyl-homoserine lactone synthase. *P Natl Acad Sci USA* **108:** 12089-12094

Clements A, Rojas JR, Trievel RC, Wang L, Berger SL, Marmorstein R (1999) Crystal structure of the histone acetyltransferase domain of the human PCAF transcriptional regulator bound to coenzyme A. *European Molecular Biology Organization* **18:** 3521-3532

Collaborative Computational Project N (1994) The Ccp4 Suite - Programs for Protein Crystallography. *Acta Crystallographica Section D: Biological Crystallography* **50:** 760-763

Congreve M, Murray CW, Blundell TL (2005) Keynote review: Structural biology and drug discovery. *Drug Discov Today* **10:** 895-907

Cook DN, Pisetsky DS, Schwartz DA (2004) Toll-like receptors in the pathogenesis of human disease. *Nat Immunol* **5:** 975-979

Cruz-Migoni A (2011) Structural studies on BPSL1549 protein, a putative determinant of pathogenicity in *Burkholderia pseudomallei* PhD Thesis, Molecular Biology and Biotechnology, University of Sheffield, Sheffield

Cruz-Migoni A, Hautbergue GM, Artymiuk PJ, Baker PJ, Bokori-Brown M, Chang CT, Dickman MJ, Essex-Lopresti A, Harding SV, Mahadi NM, Marshall LE, Mobbs GW, Mohamed R, Nathan S, Ngugi SA, Ong C, Ooi WF, Partridge LJ, Phillips HL, Raih MF, Ruzheinikov S, Sarkar-Tyson M, Sedelnikova SE, Smither SJ, Tan P, Titball RW, Wilson SA, Rice DW (2011) A *Burkholderia pseudomallei* toxin inhibits helicase activity of translation factor eIF4A. *Science* **334:** 821-824

Currie B, Ward L, Cheng A (2010) The Epidemiology and Clinical Spectrum of Melioidosis: 540 Cases from the 20 Year Darwin Prospective Study. *Plos Neglected Tropical Diseases* **4:** e900

Currie BJ, Dance DA, Cheng AC (2008) The global distribution of *Burkholderia pseudomallei* and melioidosis: an update. *T Roy Soc Trop Med H* **102 Suppl 1:** S1-4

Dauter Z (1999) Data-collection strategies. *Acta Crystallographica Section D: Biological Crystallography* **55:** 1703-1717

DeShazer D, Brett PJ, Woods DE (1998) The type II O-antigenic polysaccharide moiety of *Burkholderia pseudomallei* lipopolysaccharide is required for serum resistance and virulence. *Molecular Microbiology* **30:** 1081-1100

Dieffenbach CW, Lowe TMJ, Dveksler GS (1993) General Concepts for Pcr Primer Design. *Pcr Meth Appl* **3:** S30-S37

Dry S, McCarthy S, Harris T (2000) Structural genomics in the biotechnology sector. *Nat Struct Biol* **7:** 946-949

Dutnall RN, Tafrov ST, Sternglanz R, Ramakrishnan V (1998) Structure of the Histone Acetyltransferase Hat1: A Paradigm for the GCN5-Related N-acetyltransferase Superfamily. *Cell* **94:** 427-438

Dyda F, Klein DC, Hickman AB (2000) GCN5-related N-acetyltransferases: Structural overview. *Annu Rev Biophys Biomol Struct* **29:** 81-103

Edwards AM, Arrowsmith CH, Christendat D, Dharamsi A, Friesen JD, Greenblatt JF, Vedadi M (2000) Protein production: feeding the crystallographers and NMR spectroscopists. *Nat Struct Biol* **structural genomics supplement:** 970-972

Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallographica Section D: Biological Crystallography* **66:** 486-501

Essex-Lopresti AE, Boddey JA, Thomas R, Smith MP, Hartley MG, Atkins T, Brown NF, Tsang CH, Peak IRA, Hill J, Beacham IR, Titball RW (2005) A type IV pilin, PilA, contributes to adherence of *Burkholderia pseudomallei* and virulence in vivo. *Infection and Immunity* **73:** 1260-1264

Evans G, Axford D, Owen RL (2011) The design of macromolecular crystallography diffraction experiments. *Acta Crystallographica Section D: Biological Crystallography* **67:** 261-270

Evans G, Pettifer RF (2001) CHOOCH: a program for deriving anomalous-scattering factors from X-ray fluorescence spectra. *J Appl Crystallogr* **34:** 82-86

Evans P (2006) Scaling and assessment of data quality. *Acta Crystallographica Section D: Biological Crystallography* **62:** 72-82

Evans PR (2011) An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallographica Section D: Biological Crystallography* **67:** 282-292

Firczuk M, Mucha A, Bochtler M (2005) Crystal Structures of Active LytM. *J Mol Biol* **354:** 578-590

Galyov EE, Brett PJ, DeShazer D (2010) Molecular Insights into *Burkholderia pseudomallei* and *Burkholderia mallei* Pathogenesis. In *Annual Review of Microbiology, Vol 64, 2010*, Gottesman S, Harwood CS (eds), Vol. 64, pp 495-517.

Gan YH (2005) Interaction between *Burkholderia pseudomallei* and the host immune response: Sleeping with the enemy? *Journal of Infectious Diseases* **192:** 1845-1850

Garman EF (2010) Radiation damage in macromolecular crystallography: what is it and why should we care? *Acta Crystallographica Section D: Biological Crystallography* **66:** 339-351

Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of Hidden Markov Models that represent all proteins of known structure. *J Mol Biol* **313:** 903-919

Gould TA, Schweizer HP, Churchill MEA (2004) Structure of the *Pseudomonas aeruginosa* acylhomoserinelactone synthase LasI. *Mol Microbiol* **53:** 1135-1146

Hawley RJ, Eitzen EM (2001) Biological weapons - A primer for microbiologists. *Annu Rev Microbiol* **55:** 235-253

He H, Ding Y, Bartlam M, Sun F, Le Y, Qin X, Tang H, Zhang R, Joachimiak A, Liu J, Zhao N, Rao Z (2003) Crystal Structure of Tabtoxin Resistance Protein Complexed with Acetyl Coenzyme A Reveals the Mechanism for b-Lactam Acetylation. *J Mol Biol* **325:** 1019-1030

Hickman AB, Namboodiri MAA, Klein DC, Dyda F (1999) The Structural Basis of Ordered Substrate Binding by Serotonin N-Acetyltransferase: Enzyme Complexat 1.8 A° Resolution with a Bisubstrate Analog. *Cell* **97:** 361-369

Hofmann K, Baron M. *Boxshade*.

Holden MTG (2009). High throughput sequencing of *Burkholderia pseudomallei*. *European Melioidosis Network Annual Meeting*; 11 September 2009; London School of Hygiene and Tropical Medicine, England.

Holden MTG, Titball RW, Peacock SJ, Cerdeno-Tarraga AM, Atkins T, Crossman LC, Pitt T, Churcher C, Mungall K, Bentley SD, Sebaihia M, Thomson NR, Bason N, Beacham IR, Brooks K, Brown KA, Brown NF, Challis GL, Cherevach I, Chillingworth T, Cronin A, Crossett B, Davis P, DeShazer D, Feltwell T, Fraser A, Hance Z, Hauser H, Holroyd S, Jagels K, Keith KE, Maddison M, Moule S, Price C, Quail MA, Rabbinowitsch E, Rutherford K, Sanders M, Simmonds M, Songsivilai S, Stevens K, Tumapa S, Vesaratchavest M, Whitehead S, Yeats C, Barrell BG, Oyston PCF, Parkhill J (2004) Genornic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. *P Natl Acad Sci USA* **101:** 14240-14245

Holm L RP (2010) Dali server: conservation mapping in 3D. *Nucl Acids Res* **38:** W545-549

Hu LI, Lima BP, Wolfe AJ (2010) Bacterial protein acetylation: the dawning of a new age. *Molecular Microbiology* **77:** 15-21

Inglis TJJ, Sagripanti JL (2006) Environmental factors that affect the survival and persistence of *Burkholderia pseudomallei*. *Appl Environ Microb* **72:** 6865-6875

Jahnke W, Widmer H (2004) Protein NMR in biomedical research. *Cellular and Molecular Life Sciences* **61:** 580-599

Jones AL, Beveridge TJ, Woods DE (1996) Intracellular survival of *Burkholderia pseudomallei*. *Infection and Immunity* **64:** 782-790

Jones S, Thornton JM (1995) Protein-protein interactions: A review of protein dimer structures. *Progress in Biophysics and Molecular Biology* **63:** 31-65

Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica section A* **32:** 922-923

Kabsch W (2010) XDS. *Acta Crystallographica Section D: Biological Crystallography* **66:** 125-132

Kang J, Lee MS, Gorenstein DG (2005) The enhancement of PCR amplification of a random sequence DNA library by DMSO and betaine: Application to in vitro combinatorial selection of aptamers. *J Biochem Biophys Methods* **64:** 147-151

Kantardjieff KA, Rupp B (2003) Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Sci* **12:** 1865-1871

Khan SN, Khan AU (2010) Role of histone acetylation in cell physiology and diseases: An update. *Clinica Chimica Acta* **411:** 1401-1411

Kohanski MA, Dwyer DJ, Collins JJ (2010) How antibiotics kill bacteria: from targets to networks. *Nature Rev Microbiol* **8:** 423-435

Krissinel E (2009) Crystal contacts as nature's docking solutions. *J Comput Chem* **31:** 133-143

Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **372:** 774-797

Krogh A, Larsson B, Heijne Gv, Sonnhammer ELL (2001) Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to complete Genomes. *J Mol Biol* **305:** 567-580

Laskowski RA, Macarthur MW, Moss DS, Thornton JM (1993) Procheck - a Program to Check the Stereochemical Quality of Protein Structures. *J Appl Crystallogr* **26:** 283-291

Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. *Nature reviews Molecular cell biology* **8:** 995-1005

Leslie AGW (2006) The integration of macromolecular diffraction data. *Acta Crystallographica Section D: Biological Crystallography* **62:** 48-57

Leslie AGW, Powell HR (2007) Processing Diffraction Data with Mosflm. In *Evolving Methods for Macromolecular Crystallography* Vol. 245, pp 41-51.

Limmathurotsakul D, Chantratita N, Teerawattanasook N, Piriyagitpaiboon K, Thanwisai A, Wuthiekanun V, Day NPJ, Cooper B, Peacock SJ (2011) Enzyme-Linked Immunosorbent Assay for the Diagnosis of Melioidosis: Better Than We Thought. *Clin Infect Dis* **52:** 1024-1028

Limmathurotsakul D, Chaowagul W, Chierakul W, Stepniewska K, Maharjan B, Wuthiekanun V, White NJ, Day NPJ, Peacock SJ (2006) Risk factors for recurrent melioidosis in northeast Thailand. *Clin Infect Dis* **43:** 979-986

Limmathurotsakul D, Wongratanacheewin S, Teerawattanasook N, Wongsuvan G, Chaisuksant S, Chetchotisakd P, Chaowagul W, Day NPJ, Peacock SJ (2010) Increasing Incidence of Human Melioidosis in Northeast Thailand. *Am J Trop Med Hyg* **82:** 1113-1117

LiPuma JJ (2007) Update on *Burkholderia* nomenclature and resistance. *Clinical Microbiology Newsletter* **29:** 65-69

Lovell SC, Davis IW, Adrendall WB, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC (2003) Structure validation by C alpha geometry: phi,psi and C beta deviation. *Proteins* **50:** 437-450

Lye SF, Abdul Murad AM, Nathan S (2010) *Burkholderia pseudomallei* genomic islands: Variation on a theme. In *VIth World Melioidosis Congress 2010*. Townsville, Australia

Maharjan B, Chantratita N, Vesaratchavest M, Cheng A, Wuthiekanun V, Chierakul W, Chaowagul W, Day NPJ, Peacock SJ (2005) Recurrent melioidosis in patients in northeast Thailand is frequently due to reinfection rather than relapse. *J Clin Microbiol* **43:** 6032-6034

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* **39:** D225-229

Matthews BW (1968) Solvent Content of Protein Crystals. *J Mol Biol* **33:** 491-497

McCoy A. (2010) Protein Crystallography with Coffee. In School BCAS (ed.).

McCoy A, Grosse-Kunstleve R, Adams P, Winn M, Storon iL, Read R (2007) Phaser crystallographic software. *J Appl Cryst* **40:** 658-674

McCoy AJ (2004) Liking likelihood. *Acta Crystallographica Section D: Biological Crystallography* **60:** 2169-2183

McGuffin L, Bryson K, Jones D (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16:** 404-405

Moore RA, Tuanyok A, Woods DE (2008) Survival of *Burkholderia pseudomallei* in water. *BMC Res Notes* **1:** 11

Murshudov GN, Skubak P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F, Vagin AA (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D: Biological Crystallography* **67:** 355-367

Musso M, Bocciardi R, Parodi S, Ravazzolo R, Ceccherini I (2006) Betaine, Dimethyl Sulfoxide, and 7-Deaza-dGTP, a Powerful Mixture for Amplification of GC-Rich DNA Sequences. *Journal of Molecular Diagnostics* **8:** 544-550

Neuwald AF, Landsman D (1997) GCN-5 related histone N-acetyltransferases belong to a diverse superfamily that includes the yeast SPTI0 protein. *Trends Biochem Sci* **22:** 154-155

Panomket P (2011) Immune response to *Burkholderia pseudomallei*. *J Med Assoc Thai* **94:** 1410-1417

Peacock SJ (2006) Melioidosis. *Curr Opin Infect Dis* **19:** 421-428

Peneff C, Mengin-Lecreulx D, Bourne Y (2001) The Crystal Structures of Apo and Complexed *Saccharomyces cerevisiae* GNA1 Shed Light on the Catalytic Mechanism of an Amino-sugar N-Acetyltransferase. *J Biol Chem* **276:** 16328-16334

Petersen T, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8:** 785-786

Pfeiffer T. (2009) Biotechnology boom raises security fears. *Reuters*.

Poole K (2002) Mechanisms of bacterial biocide and antibiotic resistance. *J Appl Microbiol* **92:** 55S-64S

Popat R, Crusz SA, Diggle SP (2008) The social behaviours of bacterial pathogens. *Brit Med Bull* **87:** 63-75

Poux AN, Marmorstein R (2003) Molecular Basis for Gcn5/PCAF Histone Acetyltransferase Selectivity for Histone and Nonhistone Substrates. *Biochemistry* **42:** 14366-14374

Pumpuang A, Chantratita N, Wikraiphat C, Saiprom N, Day NPJ, Peacock SJ, Wuthiekanun V (2011) Survival of *Burkholderia pseudomallei* in distilled water for 16 years. *T Roy Soc Trop Med H* **105:** 598-600

Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7:** 95-99

Ramadurai L, Jayaswal RK (1997) Molecular Cloning, Sequencing, and Expression of lytM, a Unique Autolytic Gene of *Staphylococcus aureus*. *J Bacteriol* **179:** 3625-3631

Ramadurai L, Lockwood KJ, Nadakavukaren MJ, Jayaswal RK (1999) Characterization of a chromosomally encoded glycylglycine endopeptidase of *Staphylococcus aureus*. *Microbiology* **145:** 801-808

Reckseidler SL, DeShazer D, Sokol PA, Woods DE (2001) Detection of bacterial virulence genes by subtractive hybridization: Identification of capsular polysaccharide of *Burkholderia pseudomallei* as a major virulence determinant. *Infection and Immunity* **69:** 34-44

Reckseidler-Zenteno SL, DeVinney R, Woods DE (2005) The capsular polysaccharide of *Burkholderia pseudomallei* contributes to survival in serum by reducing complement factor C3b deposition. *Infection and Immunity* **73:** 1106-1115

Redfern O, Harrison A, Dallman T, Pearl F, Orengo C (2007) CATHEDRAL: A Fast and Effective Algorithm to Predict Folds and Domain Boundaries from Multidomain Protein Structures. *PLoS Computational Biology* **3:** e232

Rhodes G (2006) *Crystallography made crystal clear [electronic resource] : a guide for users of macromolecular models*, Amsterdam: Boston: Amsterdam : Elsevier/Academic Press, c2006.

Rojas JR, Trievel RC, Zhou J, Mo Y, Li X, Berger SL, Allis CD, Marmorstein R (1999) Structure of Tetrahymena GCN5 bound to coenzyme A and a histone H3 peptide. *Nature* **401:** 93-98

Rotz LD, Khan AS, Lillibridge SR, Ostroff SM, Hughes JM (2002) Public health assessment of potential biological terrorism agents. *Emerg Infect Dis* **8:** 225-230

Rupp B (2010) *Biomolecular crystallography : principles, practice, and application to structural biology*, New York: New York : Garland, c2010.

Sagripanti JL, Levy A, Robertson J, Merritt A, Inglis TJJ (2009) Inactivation of Virulent *Burkholderia pseudomallei* by Sunlight. *Photochem Photobiol* **85:** 978-986

Scheibner KA, Angelis JD, Burley SK, Cole PA (2002) Investigation of the Roles of Catalytic Residues in Serotonin N-Acetyltransferase. *J Biol Chem* **277:** 18118-18126

Schell MA, Ulrich RL, Ribot WJ, Brueggemann EE, Hines HB, Chen D, Lipscomb L, Kim HS, Mrazek J, Nierman WC, DeShazer D (2007) Type VI secretion is a major virulence determinant in *Burkholderia mallei*. *Molecular Microbiology* **64:** 1466-1485

Schrodinger, LLC. (2010) The PyMOL Molecular Graphics System, Version 1.3r1.

Shalom G, Shaw JG, Thomas MS (2007) In vivo expression technology identifies a type VI secretion system locus in *Burkholderia pseudomallei* that is induced upon invasion of macrophages. *Microbiology-Sgm* **153:** 2689-2699

Shrivastava S, Mande SS (2008) Identification and Functional Characterization of Gene Components of Type VI Secretion System in Bacterial Genomes. *Plos One* **3:** e2955

Siehl DL, Castle LA, Gorton R, Keenan RJ (2007) The Molecular Basis of Glyphosate Resistance by an Optimised Microbial Acetyltransferase. *J Biol Chem* **282:** 11446-11455

Sim SH, Yu Y, Lin CH, Karuturi RKM, Wuthiekanun V, Tuanyok A, Chua HH, Ong C, Paramalingam SS, Tan G, Tang L, Lau G, Ooi EE, Woods D, Feil E, Peacock SJ, Tan P (2008) The Core and Accessory Genomes of *Burkholderia pseudomallei*: Implications for Human Melioidosis. *Plos Pathog* **4**: e1000178

Smyth MS, Martin JHJ (2000) X-Ray crystallography. *J Clin Pathol-Mol Pa* **53:** 8-14

Sternglanz R, Schindelin H (1999) Structure and mechanism of action of the histone acetyltransferase Gcn5 and similarity to other N-acetyltransferases. *P Natl Acad Sci USA* **96:** 8807-8808

Stevens MP, Galyov EE (2004) Exploitation of host cells by *Burkholderia pseudomallei*. *International Journal of Medical Microbiology* **293:** 549-555

Stevens MP, Wood MW, Taylor LA, Monaghan P, Hawes P, Jones PW, Wallis TS, Galyov EE (2002) An Inv/Mxi-Spa-like type III protein secretion system in *Burkholderia pseudomallei* modulates intracellular behaviour of the pathogen. *Molecular Microbiology* **46:** 649-659

Stockman BJ, Dalvit C (2002) NMR screening techniques in drug discovery and drug design. *Progress in Nuclear Magnetic Resonance Spectroscopy* **41:** 187-231

Stone R (2007) Infectious disease - Racing to defuse a bacterial time bomb. *Science* **317:** 1022-1024

Su YC, Wan KL, Mohamed R, Nathan S (2008) A genome level survey of *Burkholderia pseudomallei* immunome expressed during human infection. *Microbes and infection / Institut Pasteur* **10:** 1335-1345

Suputtamongkol Y, Chaowagul W, Chetchotisakd P, Lertpatanasuwun N, Intaranongpai S, Ruchutrakool T, Budhsarawong D, Mootsikapun P, Wuthiekanun V, Teerawatasook N, Lulitanond A (1999) Risk Factors for Melioidosis and Bacteremic Melioidosis. *Clin Infect Dis* **29:** 408-413

Taylor GL (2010) Experimental phasing and radiation damage Introduction to phasing. *Acta Crystallographica Section D: Biological Crystallography* **66:** 325-338

Terwilliger TC, Stuart D, Yokoyama S (2009) Lessons from Structural Genomics. *Annu Rev Biophys* **38:** 371-383

Thibault FM, Hernandez E, Vidal DR, Girardet M, Cavallo JD (2004) Antibiotic susceptibility of 65 isolates of *Burkholderia pseudomallei* and *Burkholderia mallei* to 35 antimicrobial agents. *J Antimicrob Chemother* **54:** 1134-1138

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673–4680

Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA (2000) From structure to function: Approaches and limitations. *Nat Struct Biol* **7:** 991-994

Trievel RC, Rojas JR, Sterner DE, Venkataramani RN, Wang L, Zhou J, Allis CD, Berger SL, Marmorstein R (1999) Crystal structure and mechanism of histone acetylation of the yeast GCN5 transcriptional coactivator. *P Natl Acad Sci USA* **96:** 8931-8936

Troisfontaines P, Cornelis GR (2005) Type III secretion: More systems than you think. *Physiology* **20:** 326-339

Tuanyok A, Leadem BR, Auerbach RK, Beckstrom-Sternberg SM, Beckstrom-Sternberg JS, Mayo M, Wuthiekanun V, Brettin TS, Nierman WC, Peacock SJ, Currie BJ, Wagner DM, Keim P (2008) Genomic islands from five strains of *Burkholderia pseudomallei*. *Bmc Genomics* **9:** 566

Ulrich RL, DeShazer D, Brueggemann EE, Hines HB, Oyston PC, Jeddeloh JA (2004) Role of quorum sensing in the pathogenicity of *Burkholderia pseudomallei*. *Journal of Medical Microbiology* **53:** 1053-1064

Vakulenko SB, Mobashery S (2003) Versatility of Aminoglycosides and Prospects for Their Future. *Clinical Microbiology Reviews* **16:** 430-450

Van Bambeke F, Balzi E, Tulkens PM (2000) Antibiotic efflux pumps. *Biochem Pharmacol* **60:** 457-470

Van den Ent F, Vinkenvleugel TMF, Ind A, West P, Veprintsev D, Nanninga N, den Blaauwen T, Löwe J (2008) Structural and mutational analysis of the cell division protein FtsQ. *Mol Microbiol* **68:** 110-123

Velapatino B, Limmathurotsakul D, Peacock SJ, Speert DP (2012) Identification of differentially expressed proteins from *Burkholderia pseudomallei* isolated during primary and relapsing melioidosis. *Microbes Infect* **14:** 335-340

Vetting MW, Bareich DC, Yu M, Blanchard JS (2008a) Crystal structure of RimI from *Salmonella typhimurium LT2*, the GNAT responsible for Na-acetylation of ribosomal protein S18. *Protein Sci* **17:** 1781-1790

Vetting MW, Carvalho LPSd, Roderick SL, Blanchard JS (2005a) A Novel Dimeric Structure of the RimL N-acetyltransferase from *Salmonella typhimurium*. *J Biol Chem* **280:** 22108-22114

Vetting MW, Carvalho LPSd, Yu M, Hegde SS, Magnet S, Roderick SL, Blanchard JS (2005b) Structure and functions of the GNAT superfamily of acetyltransferases. *Archives of Biochemistry and Biophysics* **433:** 212-226

Vetting MW, Hegde SS, Javid-Majd F, Blanchard JS, Roderick SL (2002) Aminoglycoside 2'-N-acetyltransferase from *Mycobacterium tuberculosis* in complex with coenzyme A and aminoglycoside substrates. *Nat Struct Biol* **9:** 653-658

Vetting MW, Magnet S, Nieves E, Roderick SL, Blanchard JS (2004) A Bacterial Acetyltransferase Capable of Regioselective N-Acetylation of Antibiotics and Histones. *Chemistry and Biology* **11:** 565-573

Vetting MW, Park CH, Hegde SS, Jacoby GA, Hooper DC, Blanchard JS (2008b) Mechanistic and Structural Analysis of Aminoglycoside N-Acetyltransferase AAC(6')-Ib and Its Bifunctional, Fluoroquinolone-Active AAC(6')-Ib-cr Variant. *Biochemistry* **47:** 9825-9835

Vetting MW, Roderick SL, Yu M, Blanchard JS (2003) Crystal structure of mycothiol synthase (Rv0819) from *Mycobacterium tuberculosis* shows structural homology to the GNAT family of N-acetyltransferases. *Protein Sci* **12:** 1954-1959

Vetting MW, Yu M, Rendle PM, Blanchard JS (2006) The Substrate-induced Conformational Change of *Mycobacterium tuberculosis* Mycothiol Synthase. *J Biol Chem* **281:** 2795-2802

Vincze T, Posfai J, Roberts RJ (2003) NEBcutter: a program to cleave DNA with restriction enzymes. *Nucleic Acids Res* **31:** 3688-3691

Warawa JM, Long D, Rosenke R, Gardner D, Gherardini FC (2009) Role for the *Burkholderia pseudomallei* Capsular Polysaccharide Encoded by the wcb Operon in Acute Disseminated Melioidosis. *Infection and Immunity* **77:** 5252-5261

Weigelt J, McBroom-Cerajewski LD, Schapira M, Zhao Y, Arrowmsmith CH (2008) Structural genomics and drug discovery: all in the family. *Curr Opin Chem Biol* **12:** 32-39

White NJ (2003) Melioidosis. *Lancet* **361:** 1715-1722

Wiersinga WJ, van der Poll T, White NJ, Day NP, Peacock SJ (2006) Melioidosis: insights into the pathogenicity of *Burkholderia pseudomallei*. *Nat Rev Microbiol* **4:** 272-282

Winsor G, Khaira B, Van Rossum T, Lo R, Whiteside M, Brinkman F (2008) The Burkholderia Genome Database: facilitating flexible queries and comparative analyses. *Bioinformatics* **24:** 2803-2804

Winter G (2010) xia2: an expert system for macromolecular crystallography data reduction. *J Appl Crystallogr* **43:** 186-190

Wolf E, Angelis JD, Khalil EM, Cole PA, Burley SK (2002) X-ray Crystallographic Studies of Serotonin N-acetyltransferase Catalysis and Inhibition. *J Mol Biol* **317**: 215-224

Wolf E, Vassilev A, Makino Y, Sali A, Nakatani Y, Burley SK (1998) Crystal Structure of a GCN5-Related N-acetyltransferase: *Serratia marcescens* Aminoglycoside 3-N-acetyltransferase. *Cell* **94:** 439-449

Wongtrakoongate P, Mongkoldhumrongkul N, Chaijan S, Kamchonwongpaisan S, Tungpradabkul S (2007) Comparative proteomic profiles and the potential markers between *Burkholderia pseudomallei* and *Burkholderia thailandensis*. *Mol Cell Probes* **21:** 81-91

Woods DE, DeShazer D, Moore RA, Brett PJ, Burtnick MN, Reckseidler SL, Senkiw MD (1999) Current studies on the pathogenesis of melioidosis. *Microbes Infect* **1:** 157-162

Wuthiekanun V, Limmathurosakul D, Chaowagul W, Chierakul W, Day NP, Peacock SJ (2010). Utility of culture in the molecular era for the diagnosis of melioidosis. *VIth World Melioidosis Congress 2010*; Townsville, Australia.

Wuthiekanun V, Smith MD, White NJ (1995) Survival of *Burkholderia pseudomallei* in the Absence of Nutrients. *T Roy Soc Trop Med H* **89:** 491-491

Wybenga-Groot LE, Draker K-a, Wright GD, Berghuis AM (1999) Crystal structure of an aminoglycoside 6'-N-acetyltransferase: defining the GCN5-related N-acetyltransferase superfamily fold. *Structure* **7:** 597-507

Yang ZR, Thomson R, McMeil P, Esnouf RM (2005) RONN: the bio-basis function neural network technique applied to the dectection of natively disordered regions in proteins *Bioinformatics* **21:** 3369-3376

Zhang R, Ou HY, Zhang CT (2004) DEG: a database of essential genes. *Nucleic Acids Res* **32:** D271-272